EDUCATIONAL RESEARCH INSTITUTE OF BRITISH COLUMBIA

Report No. 84:5

THE EFFECTS OF SCORING METHOD, TOPIC, AND MODE ON GRADE 12 STUDENTS' WRITING SCORES

N. Carlman

REPORTS

REPORT NUMBER: 84:5

TITLE: THE EFFECTS OF SCORING METHOD, TOPIC, AND MODE ON GRADE 12 STUDENTS' WRITING SCORES

GRANT NUMBER: DG 334

AUTHOR(S): Nancy Carlman

PROJECT DIRECTOR: (if different from author)

PROJECT SITE OR SAMPLE: S.D. No. 39 (Vancouver)

This report is based on a study or project funded by the ERIBC Grants Program. This report may be reproduced in whole or in part, but the author(s) and ERIBC must be credited.

ERIBC takes no responsibility for the work as it was carried out, or the conclusions drawn by the author(s).

This report is available on a loan basis (in B.C. only) or a purchase basis at the cost of $12.70 from the

EDUCATIONAL RESEARCH INSTITUTE OF BRITISH COLUMBIA
Suite 701 - 601 West Broadway
Vancouver, B.C. V5Z 4C2
Telephone: (604) 873-3801

THE EFFECTS OF SCORING METHOD, TOPIC, AND MODE ON GRADE 12 STUDENTS' WRITING SCORES

by

Nancy Carlman

A. B. (with distinction), University of Rochester, 1959

M. A. T., Wesleyan University, 1961

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

by

Special Arrangements

Nancy Carlman 1984

SIMON FRASER UNIVERSITY

April 1984

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without permission of the author.

APPROVAL

Name: Nancy Carlman

Degree: Doctor of Philosophy

Title of Thesis: The Effects of Scoring Method, Topic, and Mode on Grade 12 Students' Writing Scores

Examining Committee

Chairman: R. W. Marx

J. Tuinman
Senior Supervisor

G. Sampson
Associate Professor

M. Crowhurst
Assistant Professor

J. Kendall
External Examiner

M. C. Flanigan
Brown Professor of
Rhetoric and Composition
University of Oklahoma, Norman
External Examiner

Date approved April 6, 1984

ABSTRACT

This study investigated two questions related to the writing of grade 12 students. Are there significant differences between global (modified holistic) scores and rhetorical effectiveness (modified primary trait) scores for the same papers? Are there significant differences between scores on papers written by the same students on topics in two different modes? The analysis was done on both group scores and individuals' scores.

The study examined four 200-350 word compositions of 50 grade 12 students in Vancouver, British Columbia. They wrote on two transactional topics and two expressive topics in counterbalanced order on four occasions in English classrooms. The papers were scored by six trained raters, three rating using a global scoring rubric, three others using two rhetorical effectiveness scoring rubrics, one rubric for each mode. Interrater reliability was .73 for rhetorical effectiveness scoring and .79 for global scoring.

Analysis by *t* tests for the group produced the following results: there were significant differences between scoring methods and between modes; there were no significant differences between topics within each mode. Analysis by chi square for individuals showed no differential effects of scoring method. However, writing in different modes and on different topics within mode did produce different scores for individuals scored by rhetorical effectiveness scoring.

The following conclusions may be drawn from this study: although holistic-type and primary trait-type scoring methods are both reliable and have useful purposes for testers and teachers, because they produce significantly different scores on the same papers of groups of students, they should not be compared without qualification. If choices of topics are given on writing tests for group purposes such as large scale assessment or research, topics should all elicit the same mode and should be equated in domain of content and distance from the writer. Writing tests for purposes such as achievement, admission, placement, or competency should require students to write on more than one topic to mitigate potential injustice to individuals.

ACKNOWLEDGEMENTS

Financial support for the project was provided through a grant from the Educational Research Institute of British Columbia.

Financial support for the doctoral program was provided through a doctoral fellowship from the Social Sciences and Humanities Research Council of Canada.

TABLE OF CONTENTS

Approval ii

Abstract iii

Acknowledgements iv

Table of Contents v

List of Tables vi

List of Figures vii

Chapter 1: Introduction 1

Chapter 2: Review of Related Research 9

Chapter 3: Design and Procedures 40

Chapter 4: Analysis and Results 55

Chapter 5: Discussion and Conclusion 75

Appendices 98

References 108

LIST OF TABLES

Table 1: Interrater Reliability of Three Raters for Each of Global and Rhetorical Effectiveness Scoring 56

Table 2: Differences between Global (G) and Rhetorical Effectiveness (R) Scoring 58

Table 3: Differences between Topics within Mode for Each Scoring Method 61

Table 4: Differences between Modes for Global and Rhetorical Effectiveness Scores 63

Table 5: Differences between Global (G) and Rhetorical Effectiveness (R) Scores for Top and Bottom 15 Scorers 65

Table 6: Differences between Topics within Mode and Scoring Method for Top and Bottom 15 Scorers ...66

Table 7: Differences between Modes for Global (G) and Rhetorical Effectiveness (R) Scores for Top and Bottom 15 Scorers 67

Table 8: Relationships between Scoring Methods for Each Mode 69

Table 9: Relationships between Topics within Mode and Scoring Method for Individuals 70

Table 10: Relationships between Modes for Global and Rhetorical Effectiveness Scores for Individuals 71

Table 11: Differences between the Scores of Female and Male Writers 72

LIST OF FIGURES

Figure 1: Frequency Distributions for Group Scores 102

Figure 2: Frequency Distributions for Top and Bottom 15 Students' Scores 103

CHAPTER 1: Introduction

The purpose of this study is to examine the effects of scoring method, topic, and mode on the assessment of the writing ability of grade twelve students. One effect results from the rating scales used for the scoring of writing tests, the other from the types of topics on direct tests of writing.

Both rating scales and topic choice are potential sources of injustice for students taking writing tests. If scores from different rating systems differ significantly, naive interpreters may draw misleading conclusions about the writing ability of groups of students. More important, generalizations drawn from group results may not be pertinent to a valid measure of individuals' performances. If the types of topics set on writing tests differ significantly and if differential response takes place, comparison of students' papers written on different topics is unfair.

Concern with students' writing ability is not new. Indeed, politicians, educators from elementary schools to universities, and the public continue to demand more and more precision in determining how many students of today's schools can "write well." Achievement, placement, equivalency, regional assessment, and competency tests proliferate. Trends cycle from direct assessment of writing (writing samples) to indirect assessment (objective tests of related skills) to a combination of the two.

Recent gains in the reliability of scoring methods and

the training of raters have made direct assessment respectable again, and measurement of whole pieces of discourse to test writing is now generally accepted. Widely administered tests ask grade 12 and 13 students (aged 17-19) to write compositions ranging in length from 300 to 700 words as all or part of the instrument.

The scoring of whole pieces of writing has been done in a variety of ways, from Pass/Fail, general impression scoring (University of British Columbia English Composition Test, 1982) to computer-assisted scoring of objective features in essays (Knapp, 1972). Intensive research on the scoring of these writing samples has indicated that interrater reliability coefficients as high as .90 may be obtained if raters are trained and if they are fairly similar in background and experience (Cooper, 1977). Two widely-used, rapid means of scoring papers, holistic scoring (Cooper, 1977; Godshalk, Swineford & Coffman, 1966) and primary trait scoring (Lloyd-Jones, 1977), have not been directly compared, however, mainly because the former is norm-referenced (Melton & McCready, 1982; Mullis, 1981), and the latter is criterion-referenced. Yet it should be important for researchers, examiners, and teachers to know how their general impression ratings of papers compare with ratings related to the purposes demanded by the writing tasks, the latter being more directly connected to the way writing is judged in the world outside educational institutions.

Although recently developed scoring systems demonstrate

that interrater reliability can be maintained at a high level, less has been done to mitigate other sources of variance in an individual's writing on different occasions. There is considerable evidence that writing by the same student will vary depending on the context of the writing (Sanders & Littlefield, 1975), audience (Crowhurst & Piche, 1979; Wilkinson, Barnsley, Hanna, & Swan, 1979), the student's interest and motivation (Evans, 1979; Lloyd-Jones, 1977; Newkirk, 1979; Sanders & Littlefield, 1975; Smith, 1980), familiarity with content-matter (Hilgers, 1982; Lloyd-Jones, 1977; Melton & McCready, 1982), mode (Crowhurst, 1980b; Dixon & Stratta, 1981; Rosen, 1969; Spandel & Stiggins, 1980; White, 1980), and topic (Evans, 1979; Gilbert, 1980; McColly, 1970; Wiseman & Wrigley, 1958). Despite this considerable evidence, many large scale tests of writing still ask for only one sample of a student's writing. Thus, no matter how reliable the scoring, these one-sample tests are unlikely to provide a fair assessment of an individual's writing ability.

In a testing situation, the context is controlled: all students take the test on the same day, at the same time, in the same room or in similar rooms; and administrators are supposed to follow identical instructions. Audience has not generally been controlled directly: of the common large-scale tests, only the National Assessment of Educational Progress indicates the audience to whom the writing should be addressed, and it does not do so for all topics. However, it seems to be assumed that students in grades 12-

13 have learned to address an anonymous adult teacher/marker on such writing tests. Hoetker (1982) even suggests that no matter what the instruction sheet says the audience is supposed to be, students know the audience is composed of "hypercritical English teachers" (p. 387). Thus audience is controlled indirectly (Melton & McCready, 1982). Although Britton, Burgess, Martin, McLeod, & Rosen (1975) call attention to the student's interest or motivation, it is so difficult to control that examiners appear to ignore this source of variability.

With few exceptions the problem of content-matter knowledge is ignored. It can be controlled to some degree by providing students with content to which they can respond in writing (Hilgers, 1982). However, adding a reading task to the writing task not only introduces an additional variable--ability to read (Freedman & Robinson, 1982)--but also requires more time for the administration of the test. The University of British Columbia English Composition Test, which does provide a 600 word prose passage as content for the writing topics, requires 2.5 hours to administer.

Another widely used method of controlling content is the use of pictures as sources of information or stimuli for the writing, such as those used by Crowhurst (1977) with students in grades six and ten and by Jeroski (1982) with junior high school students. It is likely that pictures would have limited value as content for writing tests for older students who are commonly asked to write on abstract as well as concrete subjects.

Even when context and audience are controlled, evidence accumulated over the past 25 years suggests that topic choices may influence how students' papers are evaluated (Rosen, 1969; Spandel & Stiggins, 1980; Wiseman & Wrigley, 1958). Yet writing tests routinely ask students not only to make choices among topics but also among types of writing, sometimes as wide choices as between writing a story and writing a definition of an abstraction (See the 1964 General Certificate of Education Ordinary Level English Language Paper, Rosen, 1969, p. 94). Perhaps because interactions between topics and types of writing are very complicated, the tendency has been to reduce the number of topics on tests. The 1964 "O" level examination mentioned above had nine topics; in 1979 the British Columbia English Placement Test had five; in 1983 the University of British Columbia English Composition Test had two; and the annual California Freshman English Equivalency Examination gives two topics in two different modes but requires students to write on both.

However, despite strong evidence that different writing tasks require different abilities (White, 1980, p. 10), it is not clear whether differences between types of writing are indeed greater than differences between topics because studies have not focused on these differences directly. Even though we know that higher levels of syntactic complexity are associated with certain types of writing than with others (Cooper & Odell, 1977, p. xi; Crowhurst, 1980b; Crowhurst & Piche, 1979; Knapp, 1972; Rosen, 1969; Watson, 1980), and even though there are theories that certain types

of writing demand more developed cognitive skills than other types (Bereiter, 1978; Dixon & Stratta, 1982a; Egan, 1979; Rosen, 1969; Wilkinson et al., 1979; Wiseman & Wrigley, 1958), there still exists little evidence on the question whether the variance in students' performance is greater between types of writing than between topics within these types. Until this question is decided, there will continue to be uncertainty about the demands and wording of topics on writing tests.

There is general agreement among composition teachers and researchers that students write differently in different "modes"--as different types of writing are commonly called in North America--(Rosen, 1969; White, 1981). However, published evidence of the extent or reliability of these differences in terms of the quality of the writing does not exist despite Braddock, Lloyd-Jones, & Schoer's 1963 admonition to researchers that they must control for mode of discourse until "more research has been done of the effect of this element on writing performance" (pp . 8-9). (See also Rosen, 1969, p. 108 and p. 138.) This study focuses directly on this question of variance in performance when students are writing in different modes and on different topics within modes.

In summary, there are two purposes of this study of twelfth grade writers: to decide if scores on students' papers derived from two different scoring systems are significantly different, and to determine if the generally accepted supposition that students are obtaining

significantly different scores when they write in different modes is true.

The study extends the scope of previous research by adapting two widely used scoring methods to make them both criterion-referenced and both on a six point scale, thus permitting direct comparison. Also, the attempt to equate topics within mode allows differences between modes and between topics within mode to be analyzed discretely.

Thus there are two main questions.

1) Are there significant differences between global (modified holistic) scores and rhetorical effectiveness (modified primary trait) scores for the same papers?

2) Are there significant differences between scores on papers written by the same students in two different modes of discourse?

If there are significant differences between these two popular types of scoring methods, then examiners must be very careful to explain to students, parents, teachers, and other interested parties just how results of writing tests should be interpreted and caution against comparison of one type of scoring with another. If there are significant differences between scores on papers written by the same students in two different modes of discourse, examiners must take care to word all writing test topics in such a way that they will all elicit the same mode of discourse and thus be comparable in demands on the individual writing the test.

Exploration of these questions will help future examiners of written language by making clear the

relationships between two widely-used scoring methods and by elucidating one aspect of the complicated interconnections of topic and mode.

The main questions in the study were examined from the perspective of the whole group, the top and bottom fifteen students, and individuals.

Chapter 2: Review of Related Research

Research about writing has grown in scope and magnitude, particularly in the past ten years (Jeroski, 1982). It includes such traditional subjects as instruction in writing and measurement of writing as well as newer areas such as processes of writing, invention in writing, cohesion, revision, peer editing, workshop writing, cooperative writing, writing using word processors, syntactic complexity, journal writing, imitation and modeling, technical writing, writing in response to literature, beginning writing, cognition and writing, reading and writing, basic writing, and writing for students for whom English is a second language.

The research related to this study is limited to research in the assessment of writing samples as whole pieces of writing. The basis of the limitation is both theoretical and practical.

Assessment, although occurring after the end of the complicated process of writing, is becoming more and more important as education authorities centralize power over curriculum and administer large scale competency tests (Melton & McCready, 1982). As will be discussed in detail below, the most valid assessments of writing are based on ratings of samples of whole pieces of discourse.

Although some treatment studies tie assessment directly to instruction, (Jeroski, 1982; Mellon, 1969), large scale tests of writing are not related to curricula. They cannot be assumed to be content-free (as Hopkins & Stanley, 1982,

erroneously assume they should be), but they use topics which their developers presume can be written on by all students in the population to be tested. For example, the single topic used for the British Columbia language assessment pilot study was written on by 3700 students in grades 8 and 12:

> You are to write a composition (a piece of writing several paragraphs long) in which you tell about the most interesting thing you have seen, heard, read, or imagined in the past few years. You should give reasons and specific examples that will help the reader to understand your topic. (Evanechko, 1976, Vol. 3, p. 61)

Because treatment studies introduce the whole area of instruction as well as assessment, they are mentioned in this review of the literature only in cases where the type of testing used in the treatment study has direct application to overall assessment.

As well, the types of writing asked for on large scale writing assessments are circumscribed. Generally these include primarily exposition and argument with occasional requests for narration (autobiography rather than fiction). Studies of specific types of writing not generally tested in large scale assessments, for example, technical writing or writing in response to literature, are not referred to in this review of research.

Students are assessed in writing at every grade level, but many large scale tests of writing are given at the end of secondary school or the beginning of post-secondary education. Therefore, this review of research does not mention studies of beginning or elementary school writers

except when such studies clarify what sorts of expectations are reasonable for grade 12 students.

Studies of cognition and reading as related to writing may well offer important insights into why students write better on certain kinds of tasks than they do on others. However, first it is important to establish that they do show significant variations in performance.

Large scale assessments apply equally to all students in a population. No consideration is given to those students who speak English as a second language or who were brought up to speak a non-standard dialect of English. Therefore, specific studies of such sub-populations are not referenced in this study.

Thus, two areas of research in the assessment of writing relate to the present study: studies of the reliability and validity of various scoring methods for direct tests of writing and research into the importance of topic and mode variations in students' scores on writing tests. Selected research related to these two areas will be reviewed in two major sections below:

1) Reliability and Validity of Scoring Methods

2) Topic and Mode Variations.

Almost all of the research in the assessment area focuses on groups of students, even some of the small studies such as one by Watson (1980) who examined the writing of 21 high school and 21 college students. However, many of the researchers explicitly mention how individuals' performances differed from the main effects of the groups

(Crowhurst, 1977; Mellon, 1975; Rosen, 1969). In a review of the literature on the testing of writing, Raymond (1982) warned that inferences from groups must not be "applied injudiciously to the evaluation of individual papers" (p. 41). Research references to individuals will be mentioned within the review of research below.

Reliability and Validity of Scoring Methods

The first area of research which relates to this study is that of assessment of writing. Although there are many reasons for assessing writing, from regular classroom marking to admission, placement, equivalency, and competency testing of large groups, the areas of research relevant to this study include those which have examined various types of scoring systems for direct tests of writing of large groups of adolescent students. "Direct" tests of writing are those which test writing per se by taking samples of students' writing and examining these samples as whole pieces of discourse. "Indirect" tests are concerned with measures of discrete features related to the writing such as correction of usage, punctuation, or spelling errors; judgments of good and bad grammar; vocabulary knowledge and the ability to use the dictionary; and syntactic complexity of sentences -- all elements which can be counted by hand or computer. Indirect tests may list items for students to correct or rewrite, or they may require writing samples. However, in both cases the focus is the related measures rather than ability to compose whole pieces of discourse.

Although indirect tests correlate .7 to .9 with quality

scores on essays written by the same students (See An Assessment of the Writing Performance of High School Seniors, 1977; Breland & Gaynor, 1979; Checketts & Christensen, 1974; Coffman, 1966; Hiller, Marcotte, & Martin, 1969; Lloyd-Jones, 1977), these tests have been criticized by those concerned with test validity (Cooper, 1977; Gilbert, 1980; Newkirk, 1979; Wrigley, Sparrow, & Inglis, 1967) as not being tests of writing but rather of certain other abilities, like editing, general verbal capacity, or level of cognitive development.

Extensive research into indirect tests of writing has been done by Page (1966,1967), his followers, Hiller, Marcotte, & Martin, (1969), Knapp (1972), and Finn (1977), all of whom used computers to do the counting. These studies show advances in the ability of researchers to program computers to make increasingly more sophisticated distinctions. However, educators remain convinced "that there are qualities in writing that cannot be measured by objective means" (Melton & McCready, 1982, p.6; see also Cooper, 1977). They prefer to ask students to produce whole pieces of discourse. Thus educators sample directly students' ability to write.

Direct measures of writing ability, that is the measurement of whole pieces of students' writing, allow assessors to scrutinize real pieces of discourse which are the result of all the complex skills involved in what is termed "writing."

Reliability

With any test of writing, examiners want to make sure that they are reliably testing an individual's ability to write. Reliability for indirect tests of writing can be achieved more easily than for direct tests of writing. Objective tests of discrete writing skills or skills related to writing can be checked for reliability through alternate forms, split-half, or test-retest methods (Bailey, 1978, p. 62; Guilford & Fruchter, 1978, p. 414).

Two types of reliability of direct tests of writing are relevant to this study: inter-topic reliability and interrater reliability. The former is closely related to validity and is likely to be lower than interrater reliablity since no two different questions could elicit exactly the same level of writing skills from any individual (Hopkins & Stanley, 1982, p. 211).

Also, scarcely any studies have attempted to equate topics. Interset correlations in the Rosen study (1969) ranged from .45 between sets 4 (one word titles) and 9 (a free choice, informative article) to .79 between sets 4 (one word titles) and 5 (description) (p. 161). Reliability coefficients were .62 and .88, respectively. Since no attempt was made to equate topics and each set of topics allowed broad choice within the set, no conclusions can be drawn about these relationships. Although direct tests of writing cannot be expected to achieve as high reliability coefficients as objective tests, care in equating topics should achieve values higher than those usually reported.

Higher reliability coefficients should make the tests more useful (Guilford & Fruchter, 1978, p. 87), more respectable in testing circles, and, most impórtant, fairer to individuals (Thompson, 1980).

Interrater reliability is a more important concern because direct assessment of writing must rely on the scoring of whole pieces of writing by fallible human judges. Fears that interrater reliability could never be comparable to the reliability of standardized tests of writing-related skills slowed adoption of direct assessment as a main method of testing writing. Large scale tests by Coffman (1966) and Checketts and Christensen (1974) indicated that it was very difficult to increase the reliability of the scoring of writing samples over that of a combination of a writing sample with an objective test. However, the decade between the Coffman study and the publication of Cooper's review of direct assessment (1977) saw a number of research studies of the testing of writing which were able to increase interrater reliabilities to acceptable levels. A later study of a large number of state and local testing programs reported interrater reliabilities ranging from .62 to .90 (Melton & McCready, 1982).

These studies and large scale tests used different types of scoring methods, different numbers of raters, and different selection criteria for raters. Some trained and monitored raters; some did not. Different methods were used to calculate interrater reliabilities. However, despite all the differences, the accumulation of evidence has been

sufficient to convince most test administrators that writing samples can be scored reliably.

Validity

There are three major types of validity relevant to the testing of writing: content validity, face validity, and predictive validity.

Although, as mentioned above, it is possible to get objective measures of aspects of writing ability which are highly correlated with raters' judgments of whole pieces of writing by the same individuals, for reasons developed below, these indirect measures are not perceived as highly desirable or sufficient.

Although indirect tests of writing are valid for the discrete skills they test (Coffman, 1966), by themselves they are not desirable because they suggest an emphasis on minor, countable aspects of writing. Educators do not want to encourage this emphasis. Indirect tests are not sufficient because they cannot deal with all aspects of whole pieces of writing. Some large testing programs such as the British Columbia Post-Secondary English Placement Test (1973-1983) use both indirect and direct tests of writing. The British Columbia test includes forty sentences for students to correct because post-secondary instructors want results of this indirect measure for screening students who need remediation and for identifying students for whom English is a second language (Anderson, 1979b).

Topics on writing tests may be construed as having

content validity since they have, by practice if not by principle, been accepted as appropriate measures for sampling the broad range of written discourse students are capable of writing (Bailey, 1978, p. 57; Cooper & Odell, 1977, p. xii; Gilbert, 1980). However, since there are no local curricular limitations on topics in broad assessment programs, content validity is threatened because not all students will be equally familiar with the content demanded by the topic (Dixon & Stratta, 1982a; Hilgers, 1982). This threat may be mitigated by pilot-testing topics among students similar to the group to be tested so that the test administrators will have a good idea whether or not the target population is sufficiently familiar with the subject matter; or content can be provided prior to the test or in the test situation itself. For example, see the University of British Columbia English Composition Test (April, 1983).

There is also concern with content validity when students are asked to write on only one topic and are thus rated on the basis of only one piece of written discourse. Content validity will be discussed in detail below under "Topic and Mode Variations" since the subjects and wording of topics are the content of the tests.

The main reason for preferring direct over indirect assessment of writing even though there may be only a slight increase in validity over objective tests (Godshalk, Swineford, and Coffman, 1966) is the face validity. Direct measures of writing ability certainly have face validity in that they do sample directly what it is that educators are

interested in testing: students' ability to write. Face validity might also be called "perceived validity." Direct assessment scores are perceived by educators as being more valid because "they are based on actual pieces of writing, on some writer's real performance" (Cooper & Odell, 1977, p. ix). Educators are concerned when there are writing skill tests which ask for no writing (Ulin, 1982). In addition, direct assessment provides a model for students and teachers. Godshalk, Swineford, and Coffman (1966) recognized this benefit in their discussion of the College Entrance Examination Board's tests of writing in the United States.

> An essay in the English Composition Test says to the student that skill in actual writing is an important outcome of instruction. It says to the teacher that the ability to answer multiple choice questions, unless accompanied by the ability to compose answers to essay questions, is not sufficient evidence of effective teaching. (p. 41)

These comments from the United States were echoed in Britain in the same year:

> Our argument here is that the practice of continuous writing ought...to be a major part of what a pupil does in his English school work and that the form of the examination given to him ought to encourage rather than discourage such practice. (Britton, Martin, & Rosen, 1966, p. 3)

For the British Certificate of Secondary Education, a composition part of the test "was included because we felt strongly that it demands qualities which can be evoked and assessed by no other test procedure" (Wrigley, Sparrow, & Inglis, 1967, p. 4).

Predictive validity is very important for writing tests

for purposes of admission and placement. If administrators are using a writing test to determine admission or placement, they need to know that they can rely on the test to help them choose students who are likely to succeed at their institutions or in their courses. For example, research conducted during the first two cycles of the British Columbia English Placement Test project "indicated that each of the three scores (Remedial, English-as-a-Second-Language, and Composition) contributed to the ability of the EPT to predict first-year English course achievement." (Anderson, 1979a, p. 3). For 1978-79, the correlation between the English Placement Test composition scores and the first year post-secondary English grades of 1533 students was .50 (The English Placement Test, 1982).

Predictive validity is also important for equivalency and competency. If writing tests are used to determine whether or not a student may be exempted from a course, administrators need to be convinced that the test can, in fact, select those students who already have the skills which they would otherwise gain in the process of taking the course. The predictive validity of the California freshman English equivalency examination is inconclusive because freshman English grading standards vary among instructors and among post-secondary institutions (Bradley, 1979).

Competency tests are not easy to judge in terms of predictive validity. Ostensibly they indicate whether or not a student has skills of a certain standard set as a criterion for graduation from an institution. For example,

to be granted an undergraduate degree at the University of British Columbia, a student must pass the English Composition Test administered by the English Department. Whether or not this standard of competency is predictive of the student's success with writing in his or her chosen career after university has not been studied.

Holistic and Primary Trait Scoring

The scoring methods relevant to this study are termed either "holistic" or "primary trait." Most adequately described by Cooper (1977), holistic scoring is a "guided procedure for sorting or ranking written pieces" (p. 3). Cooper lists several types of such overall evaluation from rapid, impressionistic reading to the use of various scales. However, Melton and McCready (1982) insist that holistic evaluation must be norm-referenced and therefore must not be used with "a printed rubric which defines or enumerates linguistic, rhetorical, or information features" (p. 8). Mullis (1981) concurs, saying that holistic scorers "concentrate upon forming an overall impression of each paper relative to the other papers they have read" (p. 3, italics in original).

This means of evaluating large groups of papers appears to be most promising, if one does not need to know what individual students' strengths and weaknesses are. Holistic evaluation is fast and, if care is taken in the selection and training of raters, shows high interrater reliability. Cooper (1977) claims that "spending no more than two minutes

on each paper, raters can achieve a scoring reliability as high as .90 for individual writers" (p. 3). Melton and McCready's 1982 survey of 39 state and local writing assessment programs reports a reliability range of .62 to .80 for holistic scoring.

Two nationwide tests in the United States, the Writing Achievement Test of the National Assessment of Educational Progress, 1969-1979 (NAEP, 1980) and the American College Test (Spandel & Stiggins, 1980) report .79 and .75 interrater reliabilities, respectively, using two trained raters and a holistic scoring method.

The second type of scoring system relevant to this study was developed and named "primary trait" by Richard Lloyd-Jones and a committee for the first Writing Achievement Tests of the National Assessment of Educational Progress in 1969. It is criterion referenced, using different criteria for each writing exercise. These criteria are based on the rhetorical demands of each mode of discourse. For example, if a student were asked to write a paper which described something, the primary trait criteria would refer to description -- how vivid it was, how appropriate for the designated audience, how clear, how detailed, how consistent in point of view--rather than to more generalized features of all papers such as "ideas" or "style" or "mechanics." Or, if the topic asked for a statement of opinion and support, scores would be based on the number and elaboration of reasons used in support (Lloyd-Jones, 1977).

If educators assume that any writer writes equally well no matter what type of writing is required, then holistic scoring is sufficient. If, however, they accept the assumption that students write significantly differently on different tasks, as observed by Braddock et al. (1963), Conry & Jeroski (1980), Rosen (1969), and White (1981), then they need rating methods which take into account the differing rhetorical demands of the topics, as does primary trait scoring. Primary trait scoring is also more informative than holistic scoring for instructional purposes, for it can define the areas in which individual students show weaknesses (Lloyd-Jones, 1977, p. 33).

Primary trait scoring has been used in the Writing Achievement Tests of the National Assessment of Educational Progress (NAEP, 1980). Interrater reliability averaged .90 (p. 8). Melton and McCready (1981) also report .90 interrater reliability for primary trait scoring used by state and local testing programs in the United States. In describing the annual California State University and Colleges Freshman English Equivalency Examinations which use primary trait scoring, White (1980) reports interrater reliabilities ranging from .59 to .90.

Research studies which have used primary trait scoring generally report lower reliabilities because they do not ask an additional rater to reconcile discrepant scores. Gilbert (1980) reported .88 reliability in her study of the reliability, validity, and time used evaluating student writing with a primary trait scoring guide. Watson (1980)

used a primary trait rating scale to judge overall quality of papers written by superior high school and college students. Before reconciliation by a third judge, she reported interrater reliability of .68; after reconciliation, it was .70.

Summary of Reliability and Validity of Scoring Methods

Provided that we sample whole pieces of writing of types which students can be reasonably expected to produce and that these pieces of writing are rated by trained scorers who have come to consensus about how to rank papers or how to use the rating scales, we can anticipate that direct assessment of writing can be valid and reliable (Cooper, 1977, pp. 20-21).

Although there continues to be some uncertainty about terminology because research and field work in direct evaluation of writing have been uncoordinated (Cooper, 1977, p. 4), there appears to be some agreement that holistic evaluation is norm-referenced and primary trait evaluation is criterion-referenced (Melton & McCready, 1982). Although these two scoring methods have been used in many studies, they have not been directly compared.

The National Assessment of Educational Progress Writing Achievement Tests use both holistic and primary trait scoring, but not on the same papers. In terms of national percentages of "good" papers (those rated 3 or 4 on a four-point scale) what data there is is inconclusive: fewer than 50% of the students wrote "good" papers on a descriptive topic holistically; the four topics judged by primary trait

scoring had "good" papers ranging from 20% (for a humorous "Grape Peeler" exercise) to 75% for a narrative. It is difficult to draw any conclusions from these statistics because they are not meant to be compared: neither the topics nor the students are the same, and the holistic scoring is norm-referenced while the primary trait scoring is criterion-referenced.

Earlier Follman and Anderson (1967) did compare five different evaluation scales used on the same papers and found that all but one, the Diederich analytical scale, correlated .9 with one another. Analytical scales are those which designate a number of features for evaluation, such as ideas, mechanics, organization, or coherence (Diederich, 1974). However, this study did not use a primary trait rating scale because it had not yet been developed, and the researchers suggest that the high reliability of their results may have been due to the homogeneity of the group of raters (p. 199). Conry & Jeroski (1980) reported estimated correlations between their Directed Writing Scales and both holistic and analytical ratings for grade 12 students' papers at -.12 and .43, respectively. The Directed Writing Scales are similar to primary trait scales in that they are developed for specific writing tasks.

Freedman (1981) found a high correlation between holistic and analytical rating scales in her study of eight compositions of two types, comparison/contrast and argument, of 64 students at four colleges. Comparing holistic and analytical scores is not the same as comparing holistic and

primary trait scores. Analytical rating scales specify some of the features which holistic scorers take into consideration implicitly, whereas primary trait scoring bases scores on the rhetorical demands of the mode of discourse of the topic. Gilbert (1980) and Watson (1980) used primary trait scoring as a means of assessing the writing quality of secondary and university students, but they did not compare it with another type of overall evaluation.

It is clear that no direct comparison of two of the most widely used scoring methods, holistic and primary trait, has been made.

Topic and Mode Variations

The other research area of concern to this study is the examination of topic and mode variation. This area is closely related to the validity of the assessment of writing since the wording of writing topics is the basis of direct writing tests. Is a writing sample written on a certain topic a valid measure of a student's writing? Rosen's definition of "topic" is comprehensive and generally accepted. "Topic" is "the complete verbalization of the writing assignment" (Rosen, 1969, p.36). Thus topic here is not the writer's topic as in applications of Aristotle's topoi, nor is it synonymous with subject. Knapp (1972), in his study of 643 eighth and eleventh graders' writing, had three subjects, "self," "school," and "society," but he had nine topics when task directions for each of the three

subjects were written in ways designed to elicit three different types of writing. Neither is topic synonymous with "mode" which refers to the structure of a piece of writing. (See below under "Mode Assignment.") Rosen's definition is used in this study with a few noted exceptions if researchers have used the term "topic" differently.

Writer and Topic Reliability

If we expect a score on a writing test to reflect accurately a writer's ability over time and across different writing tasks, we need some evidence that the same writer would achieve a similar score on a similar test taken at a different time. Because testing costs so much, almost all writing tests are administered in one session. Most ask for only one piece of continuous writing, necessarily, therefore, on one topic (Melton & McCready, 1982). Even when more than one piece of writing is required, writer and topic variables are confounded.

> A number of researches [sic] have been carried out in which candidates have been asked to write two or more essays, but none of these has yet made a satisfactory distinction between writer reliability and assignment reliability; that is to say, variations in performance from one essay to another have not been allocated to (a) those arising by nature of the difference of the assignment (the topic, mode of discourse, the time allowed, etc.) and (b) those arising when the nature of the assignment has been kept constant. (Britton, Martin, & Rosen, 1966, p. 7)

Even if we could ignore the boredom factor and give the same students the same writing task twice as a test/retest reliability check, those same students would not produce

equivalent pieces of writing. All would be affected by maturity, personal, internal writer variables, and the fact that they would compose different pieces of writing, different both in broad aspects like thesis statements and in specific aspects like word choice. Therefore, care must be taken in the setting and grouping of topics on any one writing test.

The Setting of Topics

To "set a topic" means to arrange a topic as material to be dealt with or to compose a question to be answered.

It would be efficient and simple if test developers could test writing by setting one topic on which all students could write the kind of writing desired with interest and vigor, thus showing the scope of their writing skills adequately. However, such is not the case.

The most comprehensive study of topics for writing assignments and tests is that of Rosen (1969) who investigated the effects of differentiated writing assignments on the performance in English composition of 50 15/16 year olds from seven English secondary schools. He looked at one set of compositions each month for ten months and at these students' papers for the Ordinary Level Examinations in English language. Interset correlations range from .45 to .79. After having broadly confirmed his main hypothesis, "that different kinds of writing constitute an important variable in the study of writing ability and create a hierarchy of difficulty" (p. 10), Rosen included a section concerned with "A Qualitative Study of the Topics

and Compositions Set by Set" (p. 262) because he had discovered that in a number of cases the topics did not elicit the kind of writing desired (p. 191).

It is generally assumed that test developers will formulate topics with reasonably specific aims in mind, that they want a sample of the student's best writing, and that they recognize that a given topic will not necessarily produce a certain kind of writing (Rosen, 1969, p. 436). Given these many constraints, most test developers also recognize that no one unprepared topic could possibly be effective as a test of all students.

> Tempting though this single topic may be..., the assumption must be resisted that it is possible to compose one meaningfully-ordered set of words which place a set of pupils on a starting line like runners in a race. Even a subject like "school" will alter their positions in relation to each other. (Rosen, 1969, p. 146)

One might assume that all students could relate to "school" as a subject for writing. However, in addition to the difficulties mentioned by Rosen above, the fact that it is one word causes other problems. First, it is so open-ended that it could elicit every type of writing from narration (supposedly easiest because students are familiar with it from their earliest schooling) to "abstract topical exposition" which Bereiter (1978) suggests imposes the greatest of cognitive demands. Second, because students find it difficult to get started with a one-word impetus, they often produce singularly uninspired prose in their attempts to write something. After investigating the essays written by his subjects on one-word essay assignments, some

concrete, like "bottles," some abstract, like "hate," Rosen (1969) commented that we might see how difficult it is to write with that sort of stimulus if we recognized that by adding an "On" before those words, we set up expectations of highly literate nineteenth century-style essays (p. 330). Not only were scores on these papers the worst of the eight sets, but also "one-word topics [essay subjects] do not seem to make room for more imagination" (p. 148).

If test developers abandon single topics, one-word or not, they open up the multiple problems of topic choice.

Any finite number of choices cannot provide a wide enough variety so that all students are equally served. Several researchers stress the importance of students' having some commitment to the writing in order for it to have any of the vitality so necessary to raise it above plodding mediocrity (Britton et al., 1975; Newkirk, 1979; Rosen, 1969; Wilkinson et al., 1979). It is also important that they have some knowledge of the content matter (Hilgers, 1982; Smith, 1980). When educators consider that students must produce some connected discourse that not only makes sense and uses conventional mechanics but also has some originality in a limited amount of time with no specific preparation and that these pieces of writing may have ramifications for their futures, then they should be able to understand just how important particular topics can be.

Test makers have attempted to satisfy the problem of interest or background knowledge by giving students choices.

However, as early as 1958, Wiseman and Wrigley's study of the writing of 137 children who were given the same essay titles from the 11+ examination twice within a four month period showed that the variance between scores on different titles was still significant even when the effect of ability was removed. Although the children were given the same list of titles, they were asked to choose a different title from the list the second time. Wiseman and Wrigley used an objective test of English skills as their criterion reference for ability. Deale (1975), in reviewing assessment practices in England, commented on the injustice of giving topic choices:

> A choice of questions complicates everything. First the children have to make the choice and it is often only too apparent that some choose the wrong ones and fail to do themselves justice; and, second, for a choice to be valid the questions should either all be of equal difficulty (which is almost impossible to ensure) or the hard ones should carry more marks (and it is equally difficult to decide beforehand how much harder one is than another). (pp. 38-39)

Research has demonstrated differences not only between topics chosen but also between topics assigned. Knapp's (1972) study of 19 objective measures in essays from 643 eighth and eleventh grade students who wrote on three subjects (self, school, and society) in three modes (narration, description, and argument) found that the effects of subjects, mode of discourse, and subject-by-mode interaction were all significant beyond the .0001 level.

When children are writing on different topics, "they are in effect doing different tests and it is one of the

most difficult tasks of all to compare the performance of one child on one test with another on a different one" (Deale, 1975, p. 39).

Explicit Topic Instructions

One way to minimize variations in types of writing elicited is to give long instructions, specifying the way the students are supposed to engage in the writing (Evans, 1979). These instuctions may refer to a series of directed writing tasks as in the British Columbia Assessment (Conry & Rodgers, 1978) or to extended instructions for the production of a single, discrete composition on one topic. Greenberg (1981) gave such extended suggestions not to minimize variation but to mitigate the cognitive demands of some of the topics for her first-year post-secondary subjects. Despite the apparent contradiction that restrictive instructions might limit the writer's ability to produce anything, topics with instructions about how the students should handle the writing appear to be more successful. Mellon (1975) emphasized the need for focused topics so that evaluators can judge a writer's ability to handle a "defined rhetorical task" (p. 36). Generally such instructions include the subject and the purpose or function of the writing. Sometimes they go further and ask the student to address a certain audience or to affect a certain attitude or tone.

Hoetker (1982) and Rosen (1969) commented on the advantages and disadvantages of explicitness in instructions. Explicit instructions help students out of

inarticulateness and make sure that the intended mode is elicited. However, explicit instructions also limit students' freedom and sometimes make the task impossible for some students by virtue of the demands (Rosen, 1969, p. 442). If topic instructions ask for specifics a student may not know about, for example, "someone you know" who has "dealt with an emergency" (Rosen, 1969, p. 105), students will avoid the task or flounder.

> The more one restricts the situation in order to define a purpose and stimulate performance of a particular kind, the greater the chances that the exercises will fall outside of respondents' experiences. (Lloyd-Jones, 1977, p. 42)

Greenberg (1981) found, in fact, that there was no significant difference between scores on papers written on the same subjects with or without suggestions for how to engage in the writing or with or without suggestions that the writer relate the subject to personal experience. However, Hoetker (1982) suggested that the changes Greenberg made in the topics were too subtle for writers to note in the tension of a test situation. Research in topic construction is still inconclusive (Appleby, 1983).

Controlling for Topic

Therefore, although topics can be so explicitly structured that they consistently elicit the kind of writing desired, by so structuring topics test developers restrict the number of students who can respond to an individual topic. This factor, coupled with the problems mentioned above in relation to single topics, leads examiners to set

more than one topic on any test. The problem, then, is to make sure that the topics are comparable, a very difficult task. "How can two or more topics offer genuine alternatives and yet be equated in the demands they make upon the pupils?" (Rosen, 1969, p. 443).

For her study of junior high school students' writing, Jeroski (1982) devised parallel forms of directed writing exercises. These involved pictures and sketches as stimuli and several tasks including letters and paragraphs of description, with audiences specified. In her discussion, Jeroski commented that "where an exercise is developed to measure one or two carefully delineated skills, alternate exercises can be provided" (p. 279). The tasks were parallel, but the extent of each was far more limited than the writing of whole compositions.

In his study of syntactic fluency, also at the junior high school level, Mellon (1969) was careful to control topic through the use of parallel forms: the task directions for a character sketch differed only in that one sketch was of a relative and the other of a non-relative. For other tasks, although he equated form, he did not equate content. Mellon's topic, "How to make a dress" is really not the same as "How to entertain Grandmother or Grandfather" (Rosen, 1969, p. 443). Mellon did show that there were greater differences in students' papers between modes than between topics within modes, but he was concerned with measures of syntactic fluency, not with the overall quality of the writing. Greenberg (1981) equated topics (one on grading

systems, one on required courses) in her study of the writing of 192 college freshmen. She reported that the two topics were not significantly different (p. 43). However, since in her study each writer wrote on only one topic, topic differences for individuals were not examined.

Mode Assignment

Widely used tests not only give choices in topics but also in the types of writing demanded by the instructions. Some traditional types are called modes of discourse, particularly in North America.

"Mode" generally refers to the form or structure of the piece of writing. Conventionally, North Americans have referred to the four Aristotelian modes: narration, description, exposition, and argument. More recently, rhetoricians have devised other ways of classifying writing. Kinneavy (1971) listed aims: expressive, persuasive , informational, scientific, exploratory, or literary. Britton (1977) called his categories functions: expressive, transactional, poetic. However, "mode of discourse" has not been operationally defined (Hoetker, 1982). In any case, no matter how one categorizes writing, the types of writing overlap. Wilkinson et al. (1979) described all four of their writing tasks for 90 children aged 7, 10, and 13 in the Crediton Project as eliciting more than one of Britton's functions. Generally, very few pieces of writing fall clearly into one category or another.

Examiners recognize that it is impossible to sample "pure" modes, and they have no desire to do so. Most

persuasive pieces of writing, for example, require a certain amount of explanation; some use description or even narration to illustrate an example. However, instructions for writing should be explicit enough to elicit types of writing which are both within the range of student skills and considered desirable for the purposes of the testing. As well, they should be appealing enough to students for them to want to write.

Although there have been no studies directly addressed to this "appeal" factor, most teachers believe "student writing is most fresh and lively when students write from their own personal experiences" (Greenberg, pp. 13-14). Topics which elicit expressive writing appear to be the most appealing since, as Lloyd-Jones (1977) suggests, "writing expressively is to some extent its own reward" (p. 47). However, examiners want students to be able to write in a variety of ways; and the older they are, the less likely they are to be asked to write expressively, at least not for "school."

Although high school writing textbooks give narrative models and suggest strategies for good narration (Guth, 1979; McMaster, 1980), students soon learn that the only writing they should take seriously is expository writing. One text does not even pay lip service to narrative writing but instead leads the reader to assume that writing <u>is</u> expository writing (Hart & Heim, 1979). Commenting on their large-scale study of British adolescents' writing, Britton et al. (1975) deplored the small percentage of expressive

writing by older students. One student in Rosen's 1969 study was explicit about the stress on expository writing she inferred from her secondary school writing instruction: when given a list of topics designed to elicit autobiographical essays, she wrote 261 words, far fewer than her average on nine other papers on different types of topics, and then added a note of complaint, saying that this type of topic was not what she was used to writing on and not what she had been led to believe would be on her "O" level examinations at the end of the school year (p. 391).

Since there is an explicit or implicit emphasis on expository writing, becoming more pronounced as students move from secondary to post-secondary education, somehow testers must find topics designed to elicit expository writing which are also appealing to students in their last year of secondary school or their first year of college or university. The best way to find such topics is to pilot test to discover not only which topics are most popular and which elicit the types of writing wanted for evaluation but also which topics produce writing which seems at least involved, if not impelled (Britton,1977).

It is not enough to make sure that the topics elicit the type of writing desired. Examiners must also make certain that the topic choices on any given test are sufficiently similar in the type of writing elicited that they can reasonably compare one student's writing on one topic with another student's writing on another. After careful analysis of the 1964 General Certificate of

Education Examination in English Language, Rosen (1969) concluded:

> There is little point in dwelling on the practice of 'O' level examiners of offering a choice of tasks which reveal even to casual inspection that they are quite different in kind and difficulty. It has been one of the findings of this study that this practice is grossly unfair. A researcher following this example is courting disaster. (p. 442)

Braddock et al., (1963) agreed: "Largely ignored by people doing research in composition, variations in mode of discourse may have more effect than variations in topic [subject] on the quality of writing" (p. 8).

White's (1973) report of the California State University and Colleges Freshman English Equivalency Examination noted that the correlation between the two topics on the examination, both of which were written on by all 4071 candidates, was only .37. "The assumption was that essays emerging from personal experience call for quite different kinds of writing than essays comparing and contrasting quotations" (p. 38).

More recently, in discussing the difficulties encountered in the results of the first National Assessment of Educational Progress in Writing, Mellon (1975) concluded that instructions must "specify purpose, mode of discourse, voice, audience in such way that the compositions of different writers may be compared with one another and judged accordingly" (p. 35).

Controlling for Mode

A few efforts have been made to control for mode. In the 1972 study referred to above, Knapp attempted to use the same three subjects across modes (narration, description, argument). His instructions, however, differed considerably for the three subjects and three modes. He did not just say "Describe your self, school, society" or "Tell a story about your self, school, society" but rather gave such elaborated instructions that his three subjects in effect became nine topics reflecting the interactions between subject and mode. In addition, students' interpretations of the instructions were various. Therefore, in that study, topic and mode were unacceptably confounded.

Crowhurst (1977) used identical pictures, such as a slide of a killer whale leaping in an aquarium pool, to elicit narration, description, and argument essays from students in grades 6 and 10. The different ages of her subjects meant that the types of essays elicited differed. Seventeen of forty sixth graders in her study did not write argument essays consistently but sometimes "recorded the conversational exchanges between two persons having a disagreement" or "recounted the story of a disagreement" (p. 62). Rosen (1969) also noted "the failure of topics in a number of cases to produce the kind of writing it was hoped to elicit" (p. 146). It is possible to determine beforehand the topics which will elicit the types of writing desired by the researcher or examining committee. If adequate pilot testing had been done for the National

Assessment of Educational Progress essay title "Going to School," the wide variations in types of compositions which resulted (from straightforward, mapliké instructions to fanciful narrative) could have been minimized (Mellon, 1975, p. 35). "Preliminary trials of proposed exercises are essential to ensure that respondents will understand the situation the designers intend" (Lloyd-Jones, 1977, p. 43).

SUMMARY

Although both holistic and primary trait methods of rating writing samples have been used extensively and refined in their application and interrater reliabilities, they have not been directly compared. Nor has any effort been made to determine how use of one or the other of these scoring methods affects individuals.

In the many studies of direct assessment of writing and on many large scale tests, topic and mode have been confounded even though most educators assume that "difference in performance by mode is well-known" (White, 1981).

The study described below is an attempt to separate topic and mode sufficiently that the variables associated with each can be analyzed discretely. If it proves true that differences in scores are greater between modes than between topics within mode, then test developers will have a clear mandate to make sure all topic choices for single sample writing tests are in the same mode.

Chapter 3: Design and Procedures

Hypotheses

The following major hypotheses were investigated in this thesis:

> Hypothesis 1: there will be no significant difference between the scores on papers rated globally and scores for the same papers rated for rhetorical effectiveness.
>
> Hypothesis 2: there will be no significant difference between the scores on papers written by the same students in two different modes of discourse.

In addition, two other questions were examined:

1. What is the relationship between topic differences and mode differences?

2. What is the relationship between whole group effects and the effects for the top 15 and bottom 15 scorers?

Analyses were performed both from the perspective of group differences and for relationships for individuals.

Design of the Study

The design of the study will be decribed in terms of the subjects, the assignments, and the procedure.

Subjects

The subjects for the study were 50 twelfth grade students randomly chosen from 110 students in six classrooms

in two secondary schools in Vancouver, British Columbia, a cosmopolitan city with approximately 600,000 people. Grade 12 students (aged 17-18) were chosen for this study because they form one of the two groups most likely to be tested on their writing ability, at least in North America. The other group includes those in the first year of post-secondary institutions (grade 13, aged 18-19). Grade 12 students are preferable to grade 13 for the purposes of this study because secondary school populations are more heterogeneous than the populations of post-secondary institutions. Thus they are more likely to show a broad range of writing abilities.

Assignments

Because it is generally agreed by composition researchers that no one writing sample is a sufficient measure of a student's ability to write (see Chapter 2), this study was designed to collect four writing samples from each student, two in each of the two widely disparate types of writing, representing both expressive and transactional functions (Britton et al., 1975).

To highlight differences in the writing of the same individuals in two different modes of discourse, researchers should choose modes which are known to differ in some aspects. It is well-known that narration and argument differ in syntactic complexity of the writing produced (Crowhurst, 1977; Rosen, 1969; Watson, 1980). It is also theorized that they differ considerably in cognitive demands

(Bereiter, 1978; Egan, 1979; Wilkinson et al., 1979).

Bossone & Larson (1980) asked 219 elementary through university teachers their opinions regarding students' problems with writing. The two problems mentioned most often were "inability to develop ideas" and "inability to organize" (p. 30)--both skills necessary for argument. Five of the seven other problems mentioned referred to mechanics. None referred to problems specific to narrative writing. British teachers tell students to avoid choosing argument titles on tests because they believe such topics are too difficult (Dixon & Stratta, 1982a). Therefore, there is a perception that students find argument more difficult than narration.

As well, narration and argument can be categorized as having different functions, "expressive" and "transactional," respectively (Britton, 1977). Expressive writing is informal, functioning as an expression of personal feelings or ideas. Transactional writing is writing "to get things done" (Britton, 1977, p. 18). Because narrative writing is so little demanded in secondary schools (at least not explicitly), I have chosen to label the description/narration topics in this study as expressive topics. Since they ask the students to describe a situation rather than a static object, the descriptions naturally lead to autobiographical narration. Continuing with the emphasis on function, I have designated the "state your opinion and support it" argument topics as transactional. Note that the topics themselves are not necessarily expressive or

transactional; in fact, all topics represent transactional writing since they are all designed to get something done, namely, to elicit certain types of writing from students. For the purposes of shorthand, the topics in this study are labeled "expressive" and "transactional" even though it is the students' writing which should be described by those functions. Thus, the topics were designed to elicit kinds of writing which differed in cognitive and syntactic demands, in structure, and in aim or function. Context and audience were kept the same, however.

Once the types of writing were established, sample topics were developed which I hoped would elicit the two types of writing explained above. It was also important that the topics be similar in content and that the instructions for each topic within mode be identical. The topics were drawn from a variety of composition text books and writing tests and modified to suit the parameters of the two different types of writing. Topics were pilot-tested twice, once with 75 grade 11 students in a large Vancouver secondary school, not one of the schools used in the main study, and later with 70 first year post-secondary students. (See Appendix B for pilot test topics.)

The topics used in the study follow.

Transactional Topic I (T1)

> Traditional sex roles are now being questioned: divorced fathers sometimes have custody of small children; women in Edmonton do heavy, outdoor work for the Department of Public Works. Some people deplore this trend; others applaud it. State your position and support it.

Transactional Topic II (T2)

> Should a private club or other private organization have the right to limit membership on the basis of sex, ethnic group, or religion? State your position and support it.

Expressive Topic I (E1)

> All of us have to face various mental, physical, or emotional challenges as we go through life. Choose one such challenge from your life and describe it, including how you handled it and how you felt about it.

Expressive Topic II (E2)

> All of us know people who make or have made our lives more meaningful. Describe your relationship with one such person and tell why you feel that that relationship was or is so important to you.

As much as possible I tried to make the types of topics similar in content requirements and wording within each mode. Both transactional topics refer to modern, North American societal customs and ask the student to "state your position and support it." Both expressive topics ask for focused autobiography with reference to the way the student feels. Thus it was more likely that differences in kinds of

writing could be analyzed with less likelihood of their being confounded by differences in topic content or instructions.

Procedure

In this study, context and audience were controlled. All papers were written in 55-minute, regular class periods in secondary school English classrooms. Students wrote on the topics in counterbalanced order (see below). The four writing days occurred during a six week period in the spring of 1982. Because both schools had rotating time schedules, the writing periods occurred at different times of the day. At my request, teachers avoided scheduling any of the writing periods for the last class on Friday afternoon.

The audience was stated, real, and the same for all topics: "three English teachers who do not know you personally but who are interested in what you think and how you write." In addition, students knew that their papers would be read, graded according to criteria requested by their teachers, and returned to them with comments. They knew they were taking part in a research project, their papers would be copied, but their identities would be protected.

The teachers administered the writing of the compositions using printed and numbered topic sheets to minimize both teacher effect and contamination of the four topics in each classroom on each writing occasion. (See examples of assignment sheets in Appendix C.) The order in which each student wrote the samples was counterblanced by

listing the topics in different orders and assigning a specific order of topics to each subject's blind number. See Appendix A for examples of counterbalancing orders.

In each writing period dictionaries were allowed, but there was no prewriting talk. Occasionally the teacher would read the topic aloud to a student who was having difficulty, but she did not explain the assignment.

As each group of essays was written, I marked them according to the criteria requested by the teachers and returned the originals to the student writers to make sure that they received both response and credit for their essays. Sanders & Littlefield (1975) indicate lack of response as one problem with the rigidly controlled essay test situation (p. 148).

Scoring

The scoring procedures were organized to make sure that whatever differences there were among individuals' papers and between scoring systems were clearly defined.

The Use of Rubrics

As mentioned in Chapter 2, holistic scoring is widely used as a reliable and fast means of scoring large numbers of papers. Primary trait scoring (Lloyd-Jones, 1977) is slower (Spandel & Stiggins, 1980, p. 24), but it relates to the rhetorical demands of the mode of discourse. Because holistic scoring is norm-referenced and primary trait scoring criterion-referenced, they cannot be directly compared.

For this study, a variation of holistic scoring, called "global" (G) scoring, was developed to be criterion-referenced. Rather than sort all the papers into piles from the best group to the least good group, scorers used general rubrics on a six point scale which could refer to both types of topics. A six point scale was chosen because even numbered scales prevent mid-point clustering (Spandel & Stiggins, 1980, p. 20) and six points give room for more variation than the more common four point scales (White, 1973).

Scorers were not instructed to make sure that they used the full range of the scale since it could well be that they might not find papers which met the criteria for one or more of the levels. Training papers came not from the sample but from those written under the same circumstances but not chosen in the random selection. Like holistic scoring, however, global scoring asks that the papers be looked at as wholes without specific concern for any particular features.

Lloyd-Jones' (1977) primary trait scoring system was also modified (See Gilbert, 1980; Watson, 1980) to make it as easily comparable to the global scoring as possible. It, too, used a criterion-referenced six point scale but with different rubrics for each of the two modes. For this study it is termed the "rhetorical effectiveness" (R) score.

Scoring Rubrics for this Study

Global Scoring --for all topics.

6: This composition will be an appropriate response to the topic instructions. Although it may have occasional faults, it will be well-organized, detailed, and generally well-written.

5: This composition will be good to very good overall but may divert in some parts from the topic or not be as clearly organized as the excellent paper or have some minor mechanical errors.

4: This composition will be well-written but superficial; or it may have some errors in organization, mechanics, consistency or style which are more serious than categories 5 and 6 above.

3: This composition will misinterpret or be inconsistent in handling the topic or use overly cliched or repetitious or general language or show some errors in organization, mechanics, or style.

2: This composition will show serious weaknesses in structure, consistency, syntax, diction, and/or development.

1: This composition will exhibit very little understanding of the topic or suggest incompetence in structure, syntax, diction.

Scoring for Rhetorical Effectiveness (two rubrics)

for Transactional Topics

6: This composition takes a position, gives two or more elaborated reasons, and takes opposing positions into consideration.

5: This composition takes a position and gives two or more elaborated reasons.

4: This composition takes a position and gives one elaborated or two or more unelaborated reasons.

3: This composition takes a position and gives one unelaborated reason. Or, it

doesn't take a position but gives elaborated reasons. (Or, it discusses the examples in the topic but does not take a position--added after the raters had their training session.)

2: This composition takes a position but gives no reasons.

1: This composition does not take a clear position on the issue.

for Expressive Topics

6: This composition describes vividly and clearly and integrates this description into the personal aspects of the paper.

5: This composition describes competently and makes connections between the description and the personal aspects of the paper.

4: This composition describes competently but does not make the connection clear to the personal aspects.

3: This composition does not describe competently or does not include any personal aspects of the event.

2: This composition neither describes competently nor includes any personal aspects of the event.

1: This composition does not appear to follow the topic instructions at all.

For all three rubrics, the numbered ranks would correspond roughly to the following evaluative words: 6-excellent, 5-very good, 4-good, 3-fair, 2-poor, 1-inadequate.

The rubrics for both scoring systems were developed before the training of raters began. One change, agreed to by the raters and the researcher, was made during the training session. It was added to the scoring rubric and is illustrated above in parentheses.

Raters

Six raters were hired to read and score the 200 papers (four from each of the 50 subjects in the random sample). Three raters were trained for the global scoring and three for the rhetorical effectiveness scoring. Deale (1975) states that three markers show reliability over two but increasing the number over three shows that the gain in reliability begins to flatten out (p. 71). Steele (1979) suggests that with three samples from each student, two raters are necessary and sufficient. And reliability figures for 2-3 raters for a large group of writing assessment projects ranged from .62-.80 (Melton & McCready, 1982).

The six raters had had from two to fifteen years of experience teaching writing and scoring essays. Four had masters' degrees in English or English Education and the other two were working toward such degrees.

Training and Scoring Sessions

The raters were randomly divided into the two scoring groups which were trained separately, using the rubrics devised for each scoring method.

The training papers were from the original data base but not from the random sample selected for analysis. I chose papers for the training which I thought deserved high, medium, and low scores. During training scorers read the rubrics, discussed them, then read and scored the papers independently. Discrepancies in scores were discussed and decisions made about how certain aspects of the papers would

be judged. The addition to the rhetorical effectiveness rubric for transactional topics (described above) was agreed to.

After 1.5 hours (global scoring) and 2 hours (rhetorical effectiveness scoring) of training, scorers separated, each one taking a set of fifty papers on the same topic. Generally accepted practice and Coffman (1971) agree that essays on the same topic should all be read at the same time. Scorers read the sets in random order and had no discussions about the papers with each other or with me for the remainder of the day.

At the beginning of the second day, I offered additional training, but the scorers preferred to continue rating papers. The total time each scorer worked varied from seven hours to ten hours for rating the 200 papers (an average of 2 -3 minutes per paper, a figure which agrees with Cooper, 1977).

The raters had no way of identifying the students except through the blind numbers, and they did not know which schools or teachers were involved. Because all papers were typed, the ratings were not affected by handwriting or neatness variations which Chase (1968) and McColly (1970) found significantly related to essays scores. All papers were typed verbatim, however, including spelling and punctuation errors.

The six raters paid attention to the topic instructions which specified the type of writing requested. For both transactional topics students were asked to "state your

position and support it." For both expressive topics students were asked to describe a situation and express their feelings about it. All three rubrics, one for global scoring and two for rhetorical effectiveness scoring, were specific in describing papers at points on the six-point scale in terms of whether or not they were appropriate responses to topic instructions.

All papers conformed to the criteria for the type of writing requested by the topic instructions except three papers from the group written on topic E2:

> All of us know people who make or have made our lives more meaningful. Describe your relationship with one such person and tell why you feel that that relationship was or is so important to you.

Two of the questioned papers described relationships with God or Jesus Christ. Although some raters questioned whether such a relationship was legitimately with a "person," the writing in each was expressive, and, on inspection, the form and development of the two papers did not differ from the other papers in the E2 set.

The third paper, from student number 335, did not describe a relationship. Instead, it gave several reasons why the writer should not have to write about personal relationships to strangers. Thus the paper was in an argument or persuasive mode rather than in the descriptive or narrative mode. Global raters generally found this a good paper despite its departure from the topic instructions. Raters 2 and 3 gave it a score of 5 whereas rater 1 gave it a 4. Rhetorical effectiveness raters were

divided. Rater 4 gave it a 5, rater 6, a 3, and rater 5, a 1.

With this exception from the 200 papers read, the remaining 199 conformed to the criteria for the mode of discourse as requested in the topic instructions.

Statistical Procedures

Before analysis of the scores to determine whether or not to reject the two null hypotheses, interrater reliabilities were calculated for both types of scoring using Ebel's (1979) formula. Confidence levels for the analysis of group and individual differences were established at <.05.

Next, analysis of the group differences between global (G) and rhetorical effectiveness (R) scores (hypothesis 1) was done through *t* tests (Guilford and Fruchter, 1978). If these had proved not significantly different, then all the scores could have been aggregated for the analysis for hypothesis 2.

Since the *t* tests for hypothesis 1 did show a significant difference between scoring systems (see Chapter 4), separate *t* tests were calculated for the different scoring systems for hypothesis 2 (differences between modes).

In addition, the scores for the top and bottom fifteen students were analyzed descriptively and calculated separately to focus more closely on how the differences affected smaller groups of certain types of writers. Correlations and chi square analyses helped focus on

individual differences.

Sex differences were noted and reported separately because they are not a focus of the study.

Chapter 4: Analysis and Results

Introduction

The results of this study are based on the scoring of 200 compositions written on four topics in two modes of discourse. The compositions were written by 50 grade 12 students (28 females, 22 males) on four occasions. They were scored by six raters using two different scoring methods.

Interrater Reliability

The six raters were randomly divided into two groups of three each. Each group of three was trained separately on one of the two scoring methods, global (G) and rhetorical effectiveness (R). The global scoring group, raters 1-3, then used only the global rating scale, while the rhetorical effectiveness group, raters 4-6, used only the rhetorical effectiveness rating scale. (See Chapter 2 for scoring rubrics.) Both groups read all 200 papers, 50 on each of the four topics from the same 50 grade 12 students.

Ebel's (1979) formula was used to calculate interrater reliability. For global scoring there was a median interrater reliability of .79; for rhetorical effectiveness scoring, it was .73. (See Table 1.)

Table 1

Interrater Reliability of Three Raters for Each of Global and Rhetorical Effectiveness Scoring

Type of Scoring	Topics T1	T2	E1	E2	Median
Global	.80	.78	.81	.78	.79
Rhetorical Effectiveness	.74	.72	.71	.74	.73

Note: T=transactional; E=expressive.

Overview of Subjects and Papers

With the exception of one 22 year old student who scored near the midpoint on a six-point scale (her mean score on all papers was 3.7), the age range of the students was from 16.5 to 18.75 with a median age of 17.5. The number of words of individual papers ranged from 216 words to 804 words.

Hypotheses

Two null hypotheses were tested in the study:

1. There will be no significant difference between the scores on papers rated globally and scores for the same papers rated for rhetorical effectiveness.

2. There will be no significant difference between the scores on papers written by the same students in two different modes of discourse.

In addition, two other questions were examined:

1. What is the relationship between topic differences and

mode differences?

2. What is the relationship between whole group effects and the effects of the top and bottom 15 scorers?

Analyses of scoring method, topic, and mode effects were done for both group and individual scores.

Scoring effects were analyzed prior to the analysis of mode effects because a finding that there were no significant differences between scoring methods would have meant that further analyses could have used a grand mean of the scores from all six raters.

Accordingly the results are reported under the following headings:

1. Scoring effects
2. Mode effects
3. Other findings.

Scoring Effects

Hypothesis 1 was investigated first; it states:

> There will be no significant difference between the scores on papers rated globally and scores for the same papers rated for rhetorical effectiveness.

Differences between means for the two types of scoring were, however, significant for three of the four topics (for both transactional topics and one of the expressive topics). Global and rhetorical effectiveness scores are listed separately with means, standard deviations, correlations, and *t* -values in Table 2.

Table 2

Differences between Global (G) and Rhetorical Effectiveness (R) Scoring

TOPICS		Means	SD	r	t	df
T1	G	3.29	1.05	.577	-4.51*	49
	R	3.86	0.85			
T2	G	3.32	1.0	.678	-6.46*	49
	R	4.01	0.86			
E1	G	3.96	0.96	.764	-1.82	49
	R	4.12	0.73			
E2	G	3.78	1.06	.704	-3.53*	49
	R	4.17	0.94			

Note: Maximum score=6; T=transactional; E=expressive.

* p =<.001

Because there were significant differences between scoring methods for three of the four topics, analysis for hypothesis 2 referring to mode effects had to be done separately for each scoring method.

Mode Effects

Hypothesis 2 referred to mode effects. It postulated:

> There will be no significant difference between the scores on papers written by the same students in two different modes of discourse.

The analysis was done in two stages, first to determine whether there were differences between the two topics within each of the two types of writing, expressive and transactional. Such analysis was necessary to avoid confounding topic and mode difference. This analysis was done through *t* tests of the differences between the means of the scores of the 50 papers on each topic. The *t* tests were done separately for the expressive topics and for the transactional topics. Separate sets of *t* tests were done for the scores from each of the two scoring methods, global and rhetorical effectiveness.

Second, the means of the two scores on expressive topics and the two scores on transactional topics were analyzed for significant differences between the two modes through the use of *t* tests. Again, separate analyses were done for each of the two scoring methods.

Differences between Topics

Use of two-tailed *t* tests indicated that there were no significant differences between transactional topics or between expressive topics for either global or rhetorical effectiveness scoring. (See Table 3.)

Table 3

Differences between Topics Within Mode for Each Scoring Method

Topics	Mean	SD	r	t	df
Global Scores					
T1	3.29	1.05	.500	-0.18	49
T2	3.32	1.0			
E1	3.96	0.85	.474	1.22	49
E2	3.78	0.86			
Rhetorical Effectiveness Scores					
T1	3.86	0.96	.117	-0.95	49
T2	4.01	1.06			
E1	4.12	0.10	.256	-0.32	49
E2	4.17	0.13			

Note: T=transactional; E=expressive.

Differences between modes

Once the statistical procedures described directly above had determined that there were no significant differences between the scores on the two topics within each mode for each scoring method, analysis of differences between modes was done. Means were calculated for the scores on the two transactional topics for each scoring method as well as for the scores on the two expressive topics for each scoring method. After the means were calculated, two-tailed *t* tests determined that the difference between modes for global scoring was significant at the .001 level, for rhetorical effectiveness scoring, at the lower, .05, level. (See Table 4.)

Table 4

Differences between Modes for
Global and Rhetorical Effectiveness Scores

	Modes	Mean	SD	r	t	df
Global Scores	T Score	3.31	0.89	.521	-4.64**	49
	E Score	3.87	0.87			
Rhetorical Effectiveness Scores	T Score	3.94	0.64	.461	-2.16*	49
	E Score	4.12	0.66			

Note: T=transactional; E=expressive.

* p =<.05
** p =<.001

Other Findings

Other findings of the study include effects for the top 15 and bottom 15 students, for individuals, and for males and females.

Analysis of Top 15 and Bottom 15 Students' Scores

To determine if the results for the whole group of 50 students might mask differential results from those who scored at the upper and lower ends of the scale, the same statistical tests described above were performed for the top 15 scorers and the bottom 15 scorers, separately. The means of the scores on the four topics ranged from 4.09 to 5.0 for the top 15 and from 2.47 to 3.5 for the bottom 15. Thus

they did not overlap.

The differences between scoring methods, which were significant for the whole group on all topics except E1, were also significant for the bottom 15 students' scores for all topics. However, for the top 15 students, the differences between scoring methods were not significant except for topic E1--a reversal of the results for the whole group. (See Table 5.)

The differences between topics were not significant for the whole group or for either the top or bottom group. (See Table 6.)

Differences between modes were significant for the top group. They were significant for the bottom group for global scoring but not for rhetorical effectiveness scoring. (See Table 7.)

Table 5

Differences between Global (G) and Rhetorical Effectiveness (R) Scores for Top and Bottom 15 Scorers

Topics	Means	SD	r	t	df
TOP					
GT1	4.11	0.87	.183	-1.60	14
RT1	4.51	0.62			
GT2	4.07	1.07	.698	-2.02	14
RT2	4.47	0.70			
GE1	4.89	0.64	.752	2.23*	14
RE1	4.58	0.82			
GE2	4.69	0.93	.668	-1.73	14
RE2	5.0	0.69			
BOTTOM					
GT1	2.47	0.76	.609	-4.55***	14
RT1	3.29	0.82			
GT2	2.51	0.60	.487	-4.27***	14
RT2	3.44	0.96			
GE1	3.13	0.82	.611	-2.65*	14
RE1	3.58	0.58			
GE2	2.89	0.85	.731	-3.52**	14
RE2	3.47	0.88			

* $p =< .05$
** $p =< .01$
*** $p =< .001$

Table 6

Differences between Topics Within Mode and Scoring Method for Top and Bottom 15 Scorers

	Topics	Means	SD	r	t	df
TOP						
	GT1	4.11	0.87	.366	0.16	14
	GT2	4.07	1.07			
	RT1	4.51	0.62	-.225	0.17	14
	RT2	4.47	0.70			
	GE1	4.89	0.64	-.022	0.68	14
	GE2	4.69	0.93			
	RE1	4.58	0.82	-.126	-1.44	14
	RE2	5.0	0.69			
BOTTOM						
	GT1	2.47	0.76	.134	-0.19	14
	GT2	2.51	0.60			
	RT1	3.29	0.82	-.298	-0.42	14
	RT2	3.44	0.96			
	GE1	3.13	0.82	-.217	0.73	14
	GE2	2.89	0.85			
	RE1	3.58	0.58	-.222	0.37	14
	RE2	3.47	0.88			

Note: G=Global, R=Rhetorical Effectiveness

Table 7

Differences between Modes for Global (G) and Rhetorical Effectiveness (R) Scores for Top and Bottom 15 Scorers

Mode and Scoring Method	Means	SD	r	t	df
TOP					
TG	4.09	0.80			
			.058	-2.85*	14
EG	4.79	0.56			
TR	4.49	0.41			
			.402	-2.30*	14
ER	4.79	0.50			
BOTTOM					
TG	2.49	0.52			
			-.014	-2.73*	14
EG	3.01	0.52			
TR	3.37	0.53			
			-.195	-0.78	14
ER	3.52	0.47			

Note: T=transactional, E=expressive

* p =<.05

Relationships for Individuals

To highlight the position of individuals within the group, chi square analysis was performed on the relationships between scoring methods, topics, and modes. Even though the differences between topics within mode were not significant for the whole group while differences between modes were significant, an analysis of correlations suggests there may have been important differential effects.

It is difficult to describe individual differences statistically with an N of 4 (4 papers for each writer). However, using an arbitrary criterion of scores 4-6=pass, scores 1-3=fail, I sorted the scores for individuals on topics within mode and scoring method for chi square analysis. For global scoring the relationships between scoring methods were significant; for rhetorical effectiveness scoring, they were not. (See Table 8; for purposes of calculating chi square for the expressive scores, a value of 1 was interpreted for the Global Pass/Rhetorical Fail cell.)

Table 8

Relationships between Scoring Methods for Each Mode

Transactional

Rhetorical

Global	F	P
P	1	8
F	25	16

χ^2 =7.35*

Expressive

Rhetorical

Global	F	P
P	0	22
F	14	14

χ^2=12.12**

Note. P=pass (scores 4-6); F=fail (scores 1-3)

* p =<.01

** p =<.001

Chi square analysis of the relationships between topics within mode and scoring method for individuals showed that there were significant relationships between both expressive and transactional topics for global scoring only. The relationships between topics within mode were not significant for rhetorical effectiveness scoring for either mode. (See Table 9.)

Table 9

Relationships between Topics within Mode and Scoring Method for Individuals

Global

Topics		T2 F	T2 P
T1	P	8	12
	F	22	8

GT χ^2=5.5**

Rhetorical Effectiveness

		T2 F	T2 P
T1	P	6	26
	F	4	14

RT χ^2=.09

Topics		E2 F	E2 P
E1	P	10	22
	F	11	7

GE χ^2=4.22*

		E2 F	E2 P
E1	P	9	31
	F	3	7

RE χ^2=.25

Note. P=pass (scores 4-6); F=fail (scores 1-3)

* p =<.05

** p =<.02

Relationships between modes for individuals were significant for global scoring (p =<.02) but not for rhetorical effectiveness scoring. (See Table 10.)

Table 10

Relationships between Modes for Global and Rhetorical Effectiveness Scores for Individuals

Global

MODES	Expressive F	Expressive P
Trans. P	2	8
Trans. F	25	15

χ^2=5.82*

Rhetorical Effectiveness

	Expressive F	Expressive P
Trans. P	5	18
Trans. F	10	17

χ^2=1.38

Note. P=pass (scores 4-6); F=fail (scores 1-3)

* p =<.02

Sex Differences

Analysis of the top 15 and bottom 15 scorers showed that 13 of the top 15 scorers were female and 11 of the bottom 15 scorers were male. Calculation of means for the whole group confirmed that females scored higher than males on all topics and for both scoring methods.

Grand means for females were 3.8 for global scoring and 4.26 for rhetorical effectiveness scoring. For males, the grand means were 3.23 for global scoring and 3.6 for rhetorical effectiveness scoring. Females scored 3.4% higher than the mean for the whole group for global scoring, 3.7% higher than the mean for the whole group for rhetorical effectiveness scoring. Males scored 5.8% lower than the mean for the whole group for global scoring, 7.3% lower than the mean for the whole group for rhetorical effectiveness scoring.

Differences between grand means for males' scores and females' scores as calculated by a *t* test were significant at the *p* =<.001 level. (See Table 11.)

Table 11

Differences between the Scores of Female and Male Writers

	N	Mean	SD	t	df
F	28	4.06	.57		
				3.5*	48
M	22	3.5	.56		

* *p* =<.001

Summary of Results

Analysis of scores on 200 compositions on three of the four topics in two modes of discourse as rated by three raters for each of the two scoring methods showed that there was a significant difference between the scores on papers rated globally and scores for the same papers rated for rhetorical effectiveness.

Analysis of scores on the two topics within mode and within scoring method revealed no significant differences whereas differences between modes within scoring methods did show statistical significance. For global scoring differences between transactional and expressive mean scores were significant at the $p =< .001$ level. For rhetorical effectiveness scoring, differences between the mean scores were significant at the $p =< .05$ level.

The same statistical analysis on the papers of the top and bottom 15 scorers led to similar conclusions with some variations. Differences between scoring methods were significant for the bottom 15 but not for the top 15. Differences between topics were not significant for either the top or bottom groups. Differences between modes were significant for the top group for both methods of scoring and for the bottom group for global scoring. They were not significant for rhetorical effectiveness scoring for the bottom group.

For individuals chi square analysis showed that an individual writing in either mode would be likely to pass or fail no matter which scoring method was used. The finding

for global scoring that topics were significantly related within both transactional and expressive modes means that an individual is likely to pass or fail both topics within either mode. However, if he is scored by rhetorical effectiveness scoring, there is no such relationship. An individual is likely to pass or fail on papers written in both transactional and expressive modes if global scoring is used. However, if judged by rhetorical effectiveness scoring, she might pass if writing in one mode but fail if writing in the other.

Analysis of sex differences showed that the mean scores for females for both scoring methods and both modes were significantly higher than for males.

Chapter 5: Discussion and Conclusion

Introduction

This study of the writing of 50 grade 12 students was conducted in Vancouver, British Columbia, in the spring of 1982. This port city of 600,000 differs from other North American cities of similar size mainly in that it has many new immigrants from non-English speaking countries. The children of non-English speaking families make up approximately 39% of the public school population. To make it possible to generalize from this study to native-speakers of English in Vancouver, students for whom English is a second language were not scored by the team of raters. Of the students in the six English classes (total 166) in the two middle class secondary schools where the study was conducted, 20% (33 individuals) began school in an environment where English was the language of instruction at the age of nine or later. Therefore, the papers of those students were withdrawn from the data before scoring.

Of the remaining 133 students, papers of 23 were withdrawn either because students had written one or more papers with fewer than 200 words when the suggested length was 250-350 words or because they had not written all four papers. Students were not aware that not all papers would be used in the study, and all students' papers were marked using teacher criteria and returned to the writers.

Topics, modes, and scoring methods used in the study are similar to those used in North America in research and large scale tests for the past 20 years. In illustrating

the trend toward redefinition of sex roles in society (topic T1), reference was made to a news item about Edmonton, Alberta, since that Canadian city would be familiar to grade 12 students in a neighboring province. With that exception, the topics could be used verbatim with other North American students.

Although I have used Britton's (1977) functions, transactional and expressive, in differentiating between the two types of writing requested from the students, the topic instructions could easily be described as referring to North American modes of discourse based on the structure of the papers rather than the function. "State your position and support it" is a type of argument, not objective argument but argument which allows personal opinion. Rosen (1969) suggests that the feat of organization needed for detached argument requires too much premeditation for a writing test context (p. 473). The expressive topics asked students to describe situations and express their feelings about them, a type of description/narration very similar to the California State University and Colleges Freshman English Equivalency Examination. It sets one topic in the "expressive/analytic mode" (White, 1980, p. 10) which asks students to describe some personal experience and speculate on its meaning to the writer.

The scoring methods used were variations on holistic and primary trait scoring. Holistic scoring has been used for many years, but only in the past few decades have test developers attempted to define it specifically. Since there

is still some controversy about whether or not holistic scoring can use rubrics of any kind and whether it must always be norm-referenced (Melton & McCready, 1982), the holistic-type scoring system used in this study was designated "global" scoring.

Primary trait scoring has been used since the second National Assessment of Educational Progress Writing Achievement Tests in 1974 (Mullis, 1981). It has been modified for a variety of research and testing purposes. Some scoring guides are specifically designed for individual topics (Lloyd-Jones, 1977); others are designed for the type of writing demanded by the writing tasks (White, 1980). In this study Lloyd-Jones' scoring guide for the "Woman's Place is in the Home" writing task (Lloyd-Jones, 1977, pp. 60-64) was adapted for both transactional topics, one of which, T1, is quite similar to the "Woman's Place" topic. (See Chapter 3.) The scoring guides for the expressive topics were adapted from the California State University and Colleges Freshman English Equivalency Examination, 1979 (White, 1980, p. 14). Both scoring guides were then equated in language and for the six-point scale. Although the term "primary trait" has been used loosely by various researchers, it is associated with Lloyd-Jones and his committee for the National Assessment of Educational Progress who developed scoring guides for each topic. Since this study used a scoring guide for each mode rather than for each topic, this modification of primary trait scoring is labeled "rhetorical effectiveness" scoring.

The results of the scoring show distributions which are negatively skewed to a moderate degree, with rhetorical effectiveness scoring showing more skewness than global scoring. (See Appendix D for frequency distributions.) This result may be explained by the fact that the papers of students for whom English is a second language and students who wrote too-short papers were not included in the study. The former are likely to make grammatical errors which are negatively related to their scores (McDaniel, 1979). Longer papers are positively associated with higher scores (Knapp, 1972; Rosen, 1969).

With this exception that papers of a small number of students who presumably would score at the low end of the scale were removed before random selection, the results are thus generalizable to the middle class, English mother tongue population of grade 12 students aged 17-18 in Vancouver, British Columbia.

The purposes of this study were to determine if there are significant differences between the scores of the same papers resulting from two types of scoring and if there are significant differences between the scores of the same students writing in two modes of discourse.

Discussion

The discussion will focus first on the effects for the whole group, then for the top and bottom 15 scorers, and then for individuals. Sex differences will be discussed briefly. And finally, limitations and implications of the study and suggestions for further research will be

presented.

Whole Group Effects

Whole group effects will be discussed under scoring effects, topic effects, and mode effects.

Scoring effects. Criterion-referenced holistic scoring and primary trait scoring have not been directly compared previously. It is likely that rhetorical effectiveness (modified primary trait) scores would tend to be higher because they are focused on the rhetorical demands of the mode of discourse of the topic. Raters are not asked to pay close attention to various features of the papers like organization, paragraph structure, coherence, or mechanics but rather to concentrate on whether or not the papers fulfill the demands of the writing exercises. Although the features mentioned above would always affect raters' judgments indirectly, it is possible, for example, for a paper with many spelling errors to still be judged effective as expressive or transactional writing.

Despite the supposition that rhetorical effectiveness scores are likely to be higher than global scores, it was difficult to predict with accuracy the direction of differences in scores.

Therefore, the following null hypothesis was tested:

> There will be no significant difference between the scores on papers rated globally and scores from the same papers rated for rhetorical effectiveness.

Indeed, the results of the scoring show that there were significant differences between the scores for the same

papers rated globally (modified holistic) and in terms of rhetorical effectiveness (modified primary trait) for all topics except E1. All differences, including E1, were in the same direction: rhetorical effectiveness scores were uniformly higher than global scores. The differences between global and rhetorical effectiveness mean scores ranged from .69 for T2 to .16 for E1. As many as 34 students whose scores differed more than .69 might be affected by differences in scoring methods, depending on where the passing level is set. On a scale in which pass=scores 4-6, fail=scores 1-3, those 34 students would have failed if scored globally and passed if scored by rhetorical effectiveness scoring. (See Table 2.)

The score mean on topic E1 was the highest for global scoring (3.96) and the second highest for rhetorical effectiveness scoring (4.12). Topic E1 asked students to describe a challenge and tell how they felt about it. These papers were by far the most impelled (Britton, 1977) of the four sets. Students wrote about emotional challenges such as deaths or divorces in the family or physical challenges like climbing mountains or winning bicycle races. Raters commented that they were moved by the papers on topic E1. The rhetorical effectiveness scores were high because raters perceived that students were fulfilling the rhetorical demands of the topic.

However, global scores were also high on this topic. It may be that students "felt some inner compulsion" (Rosen, 1969, p. 98) to write and were thus somehow able to write

better papers overall than they did on other topics. Rosen's study showed that the pooled assessment marks and the exam marks were most different for what he called "discussion" topics (p. 110). In this study the marks were most different for the transactional topics and not significantly different for topic E1, an expressive topic.

Topic effects. Many educators know that students write differently on different topics. However, most of the studies or test results refer to differences not only between the subjects of the writing but between modes of discourse as well (Crowhurst, 1977; Rosen, 1969; Watson, 1980; White, 1980; Wiseman & Wrigley, 1958). Because topics are rarely equated in both subject and mode, there is not much evidence about whether differences between topics are due to the subjects or to other factors such as mode or function caused by differences in instructions. In this study, both transactional topics used the same instructions, "State your position and support it" to make as certain as possible that the mode requested was the same for each topic. The content of each topic was what differed. One was about sex roles, the other about private clubs, both societal issues.

Although both expressive topics did not have exactly the same wording for the instructions, both used the key words "describe" and "why you feel/how you felt". The subjects differed in that one topic was about a challenge, the other about a relationship, both presumably within the personal experience of each student.

In this study, therefore, for topics equated as much as is practical in mode of discourse, function, wording of instructions, and distance of the subject matter from the writer, there were no significant differences between scores on topics within mode. Mellon (1969), who also equated topics, was doing an eight month treatment study of the effects of sentence-combining exercises on syntactic complexity. He found significant increases in syntactic complexity in the papers of the students in his experimental group when they wrote on the second of two equated topics. However, the scores of the small portion (8%) of the papers rated for overall quality showed that the control group achieved higher scores after the eight month interval than did the experimental group. In the Mellon study it is difficult to differentiate among the variables: the students' maturity over eight months, teacher differences, differences in treatments, and topic variability. Other studies of overall quality have described equated topics only to the extent that they ask for the same mode of writing (Dixon & Stratta, 1982a; Rosen, 1969). The results of the present study that there were no significant differences in overall quality scores between topics within mode were the same for both scoring methods and for the whole group as well as for the top and bottom 15 scorers. However, analysis of individuals' scores showed that individuals did not achieve similar scores even on equated topics within mode. See the discussion under "Relationships for Individuals" below.

Mode effects. As mentioned under "Topic Effects" above, there are many anecdotes in the literature about differences between individual students' writing on different occasions and on different topics. In all of these studies except Mellon's (1969), topics within mode were not equated sufficiently for the mode effect to be differentiated. In the present study, however, topics were equated not only in mode but also in content domain and distance from the writer as much as was practical.

Yet, there were significant differences between modes (between the mean scores on transactional topics and on expressive topics) for both scoring methods. In all cases scores on expressive topics were significantly higher than those on transactional topics.

The fact that the statistically significant differences in this study are both in the same direction, that is, that scores on expressive topics are significantly higher than on transactional topics, may be related to the literature on syntactic complexity. Watson (1980) suggests that "In writing longer T-units [for argument topics], students take the risk of producing more syntactic anomalies than they would on a different kind of writing assignment" (p. 205). There is also considerable theory that argument topics are more difficult than expressive topics, that the ability to write argument emerges later on each person's individual schedule of cognitive development than the ability to write expressively (Bereiter, 1978; Dixon & Stratta, 1982b; Egan, 1979; Rosen, 1969; Wilkinson et al., 1979). One study which

runs counter to this trend is that of Quellmalz et al. (1982) who found that eleventh and twelfth grade students' scores on narrative essays were significantly lower than their scores on expository essays. However, these researchers do not include their topics in their report, and there may be some problem with their rating rubrics which appeared more applicable to expository writing than to narrative writing. For example, one of their rubrics mentioned "support," a term usually used as evidence for a position rather than as an aspect of narration.

Findings for the Top and Bottom 15 Students

In general the results for the whole group are similar to those for the papers of the 15 students judged to fall in each of the top and bottom groups. The top 15 are of interest because they are most likely to enter post-secondary institutions; the bottom 15 are of interest because they are students especially challenging to teachers of writing. Note that the grouping of students into top and bottom groups refers only to their mean scores on the four papers they wrote for this study. One cannot assume that these scores are related to their overall writing ability or performance in English. Some discussion of the areas in which these top and bottom groups differ from the whole group is necessary.

Differences between scoring methods. With one exception, differences between scoring methods showed the same results for the bottom 15 scorers as for the whole group: they were significant for topics T1, T2, and E2. In

contrast to the whole group, the difference between scoring methods for the bottom group was also significant for topic E1, although at the .05 level rather than the higher levels of significance for other topics. For the whole group topic E1 appeared to produce the most involved, moving writing, as it did for the bottom group as well. The mean score of the bottom group for global scoring for E1 was 3.13, for rhetorical effectiveness scoring 3.57, both scores higher than the three others for the same scoring method.

For the top fifteen students, however, the results were the reverse. Only E1 scores were significantly different for the two scoring methods ($p =< .05$); the other differences were not significant. It is clear from other studies that students who score highest write good papers both in general (See An Assessment..., 1977, for similar findings for 4000 grade 12 students in California) and also in terms of the rhetorical demands of the topics. Therefore, for these students, the choice of scoring method, as far as overall rating is concerned, is less important.

Differences between topics. Differences in scores between topics within mode were not significant for either the top 15 students or the bottom 15 students. Nor were they significant for the whole group. It is therefore reasonable to assume that equating topics is equally effective for all groups of students whatever their writing abilities.

Differences between modes. Differences in scores between modes show similar results to those for the whole

group with one exception: the difference in scores between modes was not significant for the bottom 15 scorers rated by rhetorical effectiveness scoring.

That the effects were significant for the top 15 scorers as well as for the whole group may indicate that even the writers judged best in this study have not developed their writing skills sufficiently to write equally well on both expressive and transactional topics. For the bottom 15 scorers, the difference between modes was significant for global scoring but not for rhetorical effectiveness scoring. It may be that the writers judged least good in the study were equally unable to fulfill the rhetorical demands of the two types of writing as specified by the rhetorical effectiveness scoring rubrics although they were able to write quite moving expressive papers if judged globally. Although the R scores for the bottom group were higher for all topics, they appeared to reach a sort of plateau at the 3.5 level, with a very narrow range of scores from 3.29 for topic T1 to 3.58 for topic E1. In contrast, the G scores ranged from 2.47 for T1 to 3.13 for E1. The ranges for the top 15 were broader: from 4.11 for T1 to 4.89 for E1 for G scoring and from 4.46 for T2 to 5.0 for E2 for R scoring. (See Table 8, Chapter 4.)

Relationships for Individuals

The results for individuals do not fall into a clear pattern. There are differences both between modes and between topics, yet scoring methods do not appear to affect individuals' scores to a very great extent.

These differences between individual and group effects support Raymond's (1982) contention that we cannot assume that group results will adequately represent individuals' scores. Even though this study has attempted to equate topics in mode of discourse, domain of knowledge needed, distance of the subject from the writer, audience, and length and wording of instructions, tests of writing which use these topics run the risk of being unjust to individuals, particularly if topics are assigned.

In this study, with an arbitrary "pass" level set at "4" after rating was complete, from 16 to 22 of the 50 individuals would have passed if assigned one topic within a mode and failed if assigned another topic within the same mode, depending on the mode and the scoring method. For examples of papers of students whose scores differed dramatically on two topics in the same mode, see Appendix F. In terms of actual measurement operations, this finding is highly important and should be given serious consideration in the context of the design of mass measurement of writing skills.

Most tests of writing, except provincial or national assessments, are tests to determine something about an individual. Will he pass or fail a course? Is she competent in English? Will he be admitted to a university? Can she be placed in a top section or exempted from a first year English course? If these determinations are made on the basis of one piece of writing, they are likely to be unfair to the individual because topic choice or assignment

may be crucial.

Sex Differences

The finding that females' scores are higher than males' scores on direct tests of writing is consistent with the results of the National Assessment of Educational Progress in 1974 and 1979, with a recent Australian study (Hansen, 1983), and with a California assessment of grade 12 students in 1977. In assessing the writing of 17 year olds in 1974, NAEP found that females scored 3% above the national average and males scored 4% below the national average (Hopkins & Stanley, 1982, p. 403). Similar results pertained for the 1979 assessment: for five topics females scored 4.8% above the national average and males scored 5.5% below the national average (NAEP, 1980, pp. 108-120). Hansen (1983) found that for the period 1973-1977 grade 12 females in Australia scored 15% higher than males on an English test composed mainly of writing. Comments on the California assessment of the writing of 4000 grade 12 students (An Assessment..., 1977) state that the average scores for females were slightly higher than for males.

In the present, much smaller study, 50 students as compared with 30,000 for NAEP, 12,000 for Hansen, and 4000 for the California assessment, females scored 3.5% above the study average while males scored 6.5% below the study average.

Limitations

It is important to remember that scoring effects, topic effects, and mode effects in this study apply only to four

specific topics and may not apply to other topics either assigned or chosen. No one topic can stand for the whole world of writing topics. Bradley (1979) concurs in his discussion of the reliability of the essay part of the California State University and Colleges Freshman English Equivalency Examination (White, 1980, p. 40). The results also apply only to the relationships between two pairs of four specific topics, not necessarily to all paired topics in the same realm of discourse. The modes used in this study were described as functions: expressive and transactional. However, they approximate the modal structures of description/narration and argument. The results are thus limited to these two types of writing chosen specifically to be different in syntactic and cognitive demands. Differences between modes may well not be significant if the modes are closer in syntactic and cognitive demands, such as between exposition and argument. It is also clear that differences in topic and mode cannot account for all variation in students' writing. Some variation may be attributed to unmeasurable internal variables such as the health, motivation, or interest of the writer. Additional variation may be attributed to differences among raters.

Another limitation is that whenever one calculates more than one t test on the same data, the multiple t tests tend to diminish the power of the statistical tests. However, the main conclusions of the research appear to be valid because the results are highly significant and the patterns

are generally consistent.

The top and bottom groups of scorers comprise only 15 students each, so the statistical tests are somewhat less powerful than those for the entire group. In addition, the groups' scores are not normally distributed. There is a pronounced leptokurtosis for both groups and both scoring methods. (See Appendix E for frequency distributions.) Nevertheless, again, the analysis of these restricted groups confirms in the main the findings for the whole group.

The same limitations mentioned above apply to individual effects as well as group effects. However, for individuals the fact that the topics were assigned rather than chosen is an important added limitation. In a test situation, as compared with a research situation, students are often given choices. Thus they would be likely to avoid topics which did not appeal to them or for which they did not have access to appropriate content or strategies. Also, pass/fail scores cannot be directly compared with scores on a six point scale. After scoring was completed, the "pass" level was set arbitrarily at "4" to facilitate the chi square analysis. Raters were therefore not aware that there was any "pass" level; they had just been asked to score the papers on a six point scale. How prior knowledge of a "pass" level would have affected their ratings was not analyzed.

Implications

These results have implications for both test administrators and teachers. They reinforce several good practices in test administration which have been developing on the basis of theory and research. Test administrators should be very clear about the objectives of the tests they are developing and administering. For general tests of writing ability, holistic scoring is reliable and the least expensive of all the subjective means of assessing writing (Cooper, 1977). However, if testers want to know whether or not students can fulfill the requirements of certain rhetorical tasks, primary trait scoring will produce more specific information. It is also reliable (Lloyd-Jones, 1977), although it takes longer to develop the scoring guides, train raters, and rate the papers.

In addition, this study suggests that interpreters of test results should be aware that holistic and primary trait scores produce significantly different absolute scores and, therefore, that changes from one type of evaluation to another over periods of time will make direct comparison of results difficult if not impossible.

Teachers can use primary trait scoring to discover what types of rhetorical tasks students can and cannot do and then aim their writing instruction toward helping students in the areas in which they most need help.

Many of the tests of writing commonly taken by grade 12 students are designed for those who are planning to attend post-secondary institutions, those whose achievement is at

the upper end of the scale in English as well as in other academic subjects. Therefore the results of the scores of the top group of 15 students are suggestive for students who plan to continue their education beyond secondary school.

Since differences in the two scoring methods, global and rhetorical effectiveness, are generally not significant for this top group, depending on the type of information test administrators want, either a holistic or a primary trait type of scoring system would be appropriate for this type of writer. Results from the top group of 15 show that to give this group a choice of modes on any one writing test would be unfair to those students who choose transactional topics. They would be at a disadvantage when scores on their papers are compared with those of students who wrote on expressive topics.

For teachers and for local testing purposes the finding that differences in scoring methods are significant for the bottom group of 15 writers writers is important. Those who score low on writing tests need to know that they can be convincing and moving in their writing, thus fulfilling the rhetorical demands of the topics, even though their "infelicities" (Emig, 1971) are impossible to ignore if they are being scored globally. And teachers using a primary trait type of scoring method can then help these students retain the effective writing in their papers as they revise them.

Because many large scale tests have attempted to make sure that a paper written by one student can be fairly

compared with a paper written by another, they have asked that every student write on the same topic, for example, the English Composition Test of the College Entrance Examination Board. However, they thus run the risk of setting a topic that some students may not be able to write about because they do not have knowledge of the content or because the topic does not otherwise involve them (Hilgers, 1982).

The results of this study indicate that test makers who are concerned only with group results (for example, for large scale assessment or research) can give groups of students choices on writing tests provided that they construct all topics to elicit the same mode of discourse and equate topics in length and wording of the instructions. The groups of students may be either large, heterogeneous groups or smaller, homogeneous groups, as shown by the results from the scores of the top and bottom 15 students. Topics must also be similar in the realm of knowledge needed to write about them and in the distance between the writer and the subject.

Although differences between similar topics may not matter for group results, the results for individuals show that choice of topic is important, particularly if rhetorical effectiveness scoring is used. In this study, both topics within each mode were assigned. In a test situation, a choice of topics within mode might allow individuals to choose topics about which they feel most confident in being able to fulfill the content requirements. In any case, teachers and test administrators must keep in

mind that individual variability always exists.

The finding that differences between modes of discourse are significant is very important for test makers and for teachers. Although it had been established previously that differences in syntactic complexity are associated with differences in mode of discourse (Crowhurst, 1980b, p. 10; Watson, 1980) and although "composition teachers and the Educational Testing Service are well aware of the difference in performance by mode" (White, 1981), this study gives us evidence that the differences in overall essay scores are significant.

To ignore these findings and routinely give choices of topics on writing tests which represent different types of writing is unfair to test candidates no matter whether a global (modified holistic) or rhetorical effectiveness (modified primary trait) scoring method is used. Since it is also unfair to require all students to write on one topic (Hilgers, 1982), another way to mitigate this source of injustice is to follow the California example (White, 1980) and ask students to write on two topics rather than just one during any one writing test administration. Having more than one sample required from each writer also increases the reliability of the test (Hopkins and Stanley, 1982, p. 211).

Classroom teachers at both secondary and post-secondary levels who set composition topics also need to be aware that they should make sure that all topics for any given class set of assignments should make the same types of rhetorical

demands. Otherwise comparison of students' papers is unfair. In a classroom situation, as well as in a test situation, students' writing marks should be based on more than one paper.

Suggestions for Further Research

Because the same sorts of writing tests given to grade 12 students are also given to students in their first year of post-secondary institutions for placement, equivalency, or competency, it would be valuable to repeat this study with grade 13 or first year post-secondary students. Since ability to write argument appears to develop with maturity, it may be that there would be less difference between expressive and transactional topics if subjects were older. If ability to write transactionally is maturational, it may be possible to facilitate this development. Treatment studies could be conducted to determine if the rate of this development can be increased.

Unprepared topics were set for this study. Whether or not the same types of differences would hold for prepared topics or topics for which some content is provided on the occasion of the writing would also be worth researching. As well, all topics in this study were assigned. A study which used several equated topics in each mode but which gave students choices would help determine whether topic differences for individuals are significant if students have the opportunity to avoid topics which do not appeal to them or for which they do not have access to appropriate content.

Conclusion

The findings of this study have relevance to the assessment of writing in several ways. The results relating to the differences in scoring methods make clear that scores based on holistic-type and primary trait-type scoring cannot be assumed to be comparable, although the differences for good writers are less significant. The results of the study of mode variation reinforce the view that groups of 17-18 year-old students can achieve significantly different scores on direct tests of writing when they are writing in different modes of discourse. They also show that differences between topics within mode are not significant for groups, a result which is useful for research and large scale assessment but not for testing individuals whose writing shows variability even between topics equated in rhetorical demands, in length and wording of instructions, in the realm of knowledge needed to answer the topic demands, and in the distance between the writer and the subject.

The findings are also important for teachers of composition. They must be sure that the topics they set for each classroom assignment are similar and elicit the same mode if they wish to compare students' scores, and they should keep in mind that they cannot base an individual's writing mark on any one paper. Teachers also need to know that grade 12 students are likely to write better on topics designed to elicit expressive writing than on those designed to elicit transactional writing. Thus they may wish to help

students improve their ability to write transactionally.

It is clear from these results that assessors of writing must be very careful both in their choice of method for scoring writing samples and in the setting and grouping of topics on writing tests. Otherwise they run the risk of being unfair to individuals who take these tests for admission, placement, equivalency, achievement, or competency.

APPENDIX A

Information for Teachers and Counterbalancing Examples

To: teachers whose classes are participating in a composition research project designed by Nancy Carlman.

Title: The Effects of Scoring Method, Topic, and Mode on Grade 12 Students' Writing Scores.

The purpose of this research is to find out if it is unfair to students to give them broad choices of topics on tests of writing ability. At the moment, the types of tests they commonly take (B. C. Achievement, Placement, Scholarship, and UBC English 100) give them choices among topics and among modes, sometimes as disparate as "An Unforgettable Character" and "TV or not TV."

I would like your students to write compositions on four topics (two expressive and two transactional) over a period of 4-6 weeks. The context of the writing should be kept as similar as possible: class periods of the same length and not at an unusual time of day or week, for example, not the last period on Friday.

I will mark these compositions for you in any way you wish, but I will also keep copies of the papers for marking in two different ways, global and for rhetorical effectiveness, by independent raters. These marks will be statistically analyzed to see if there are any significant differences in individuals' scores on topics in different modes.

You may have copies of a summary of the report, if you wish.

Examples of the way topics will be counterbalanced:

Student	Order of Topics
101	T1, T2, E1, E2
102	E2, T1, T2, E1
103	E1, E2, T1, T2
104	T2, E1, E2, T1
107	T2, E2, T1, E1
123	E2, E1, T1, T2

Note: T=transactional; E=expressive.

APPENDIX B

Pilot Test Topics

First Pilot Test, April, 1981

Expressive Topics: Please choose one.

1. With the flooding of Canada by U. S. movies, music, TV programs, textbooks, fashions, and fast-food outlets, there are few recognizable Canadian traits. Describe some feature of Canadian life which you feel helps make a distinction between Canadians and Americans.

2. Many sections of cities are known for an individual structure or area, like a park, a certain store, an unusual tree, a restaurant, a factory, or an ethnic community centre. Describe one feature in your neighborhood which would typify it for an outsider.

3. Think back to one of the earliest memories which made an impression on you. Describe the memory, including your place in it.

Transactional Topics: Please choose one.

1. Traditional sex roles are now being questioned: divorced fathers sometimes have custody of small children; women in Edmonton do heavy, outdoor work for the Department of Public Works. Some people deplore this trend; others applaud it. State your position and support it.

2. Should a private club or other private organization have the right to limit membership on the basis of sex, ethnic group, or religion? State your position and support it.

3. Canadian culture has often been described as a "mosaic" rather than the "melting pot" image commonly applied to the U. S. How sensible is it for ethnic groups in Canada to try to preserve their own cultural identities? Or should they assimilate into the mainstream of Canadian life? State your position and support it.

Second Pilot Test, January, 1982

Topics: Please Choose One.

1. Think back to one of the earliest memories which made an impression on you. Describe the memory, including how you felt about it.

2. All of us have to face various mental, physical, or emotional challenges as we go through life. Choose one such challenge from your life and describe it, including how you handled it and how you felt about it.

3. All of us know people who make or have made our lives more meaningful. Describe your relationship with one such person and tell why you feel that that relationship was or is so important to you.

APPENDIX C

Sample Assignment Sheet

student number 329

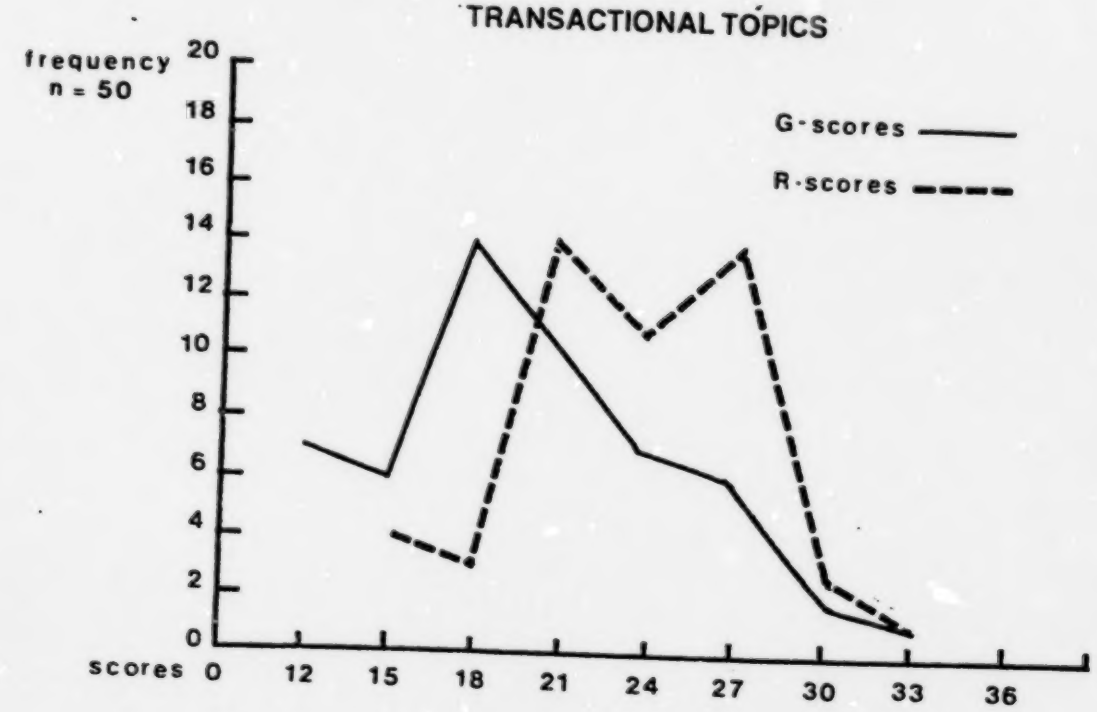
Context: regular, in-class composition.
Suggested length: 250-350 words.
Audience: three English teachers who do not know you personally but who are interested in what you think and how your write.
Please put your student number (the number indicated above) on the top right hand corner of each sheet of your composition.

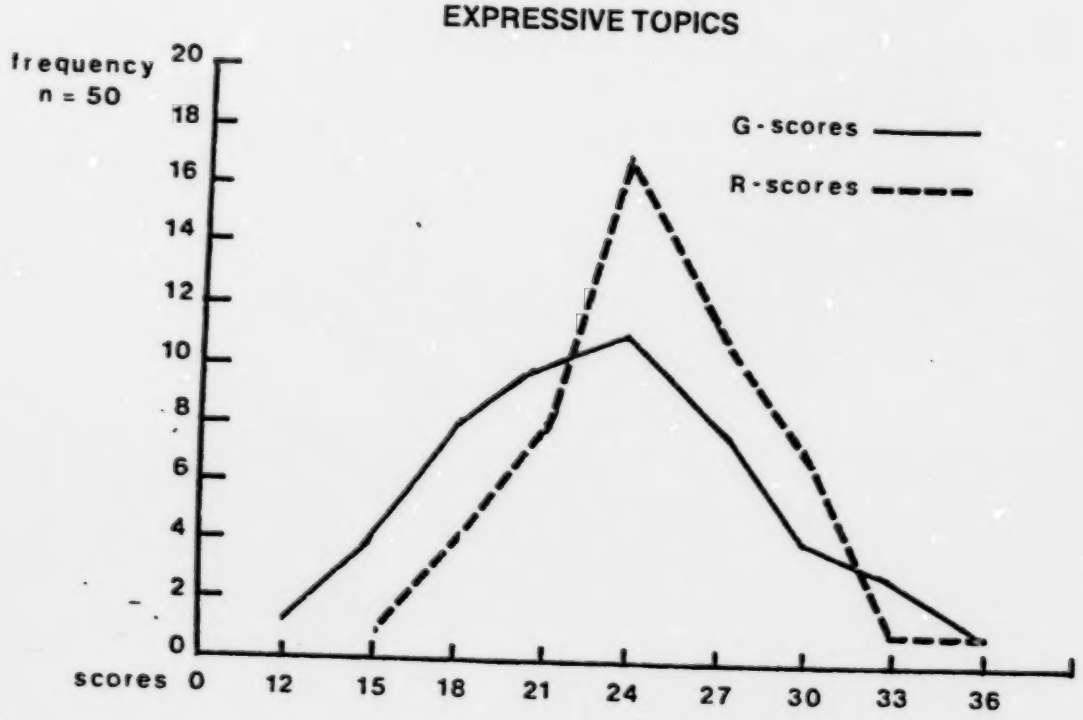
Today's Topic: T1 order of writing 3

Traditional sex roles are now being questioned: divorced fathers sometimes have custody of small children; women in Edmonton do heavy, outdoor work for the department of public works. Some people deplore this trend; others applaud it. State your position and support it.

APPENDIX D

Figure 1: Frequency Distributions for Group Scores

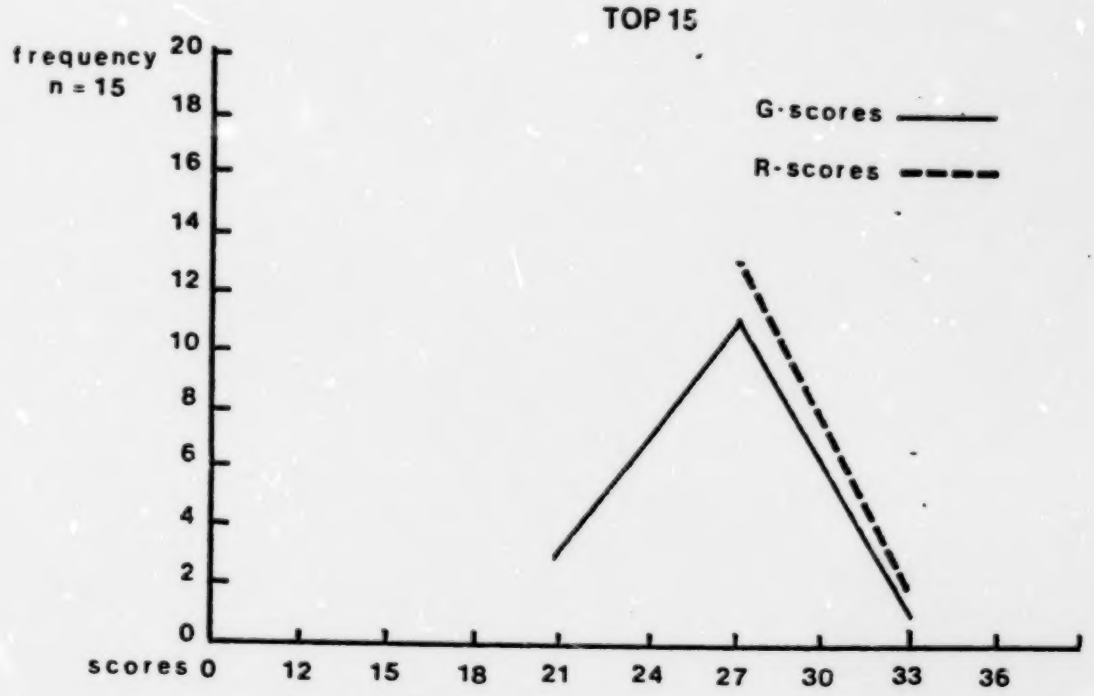


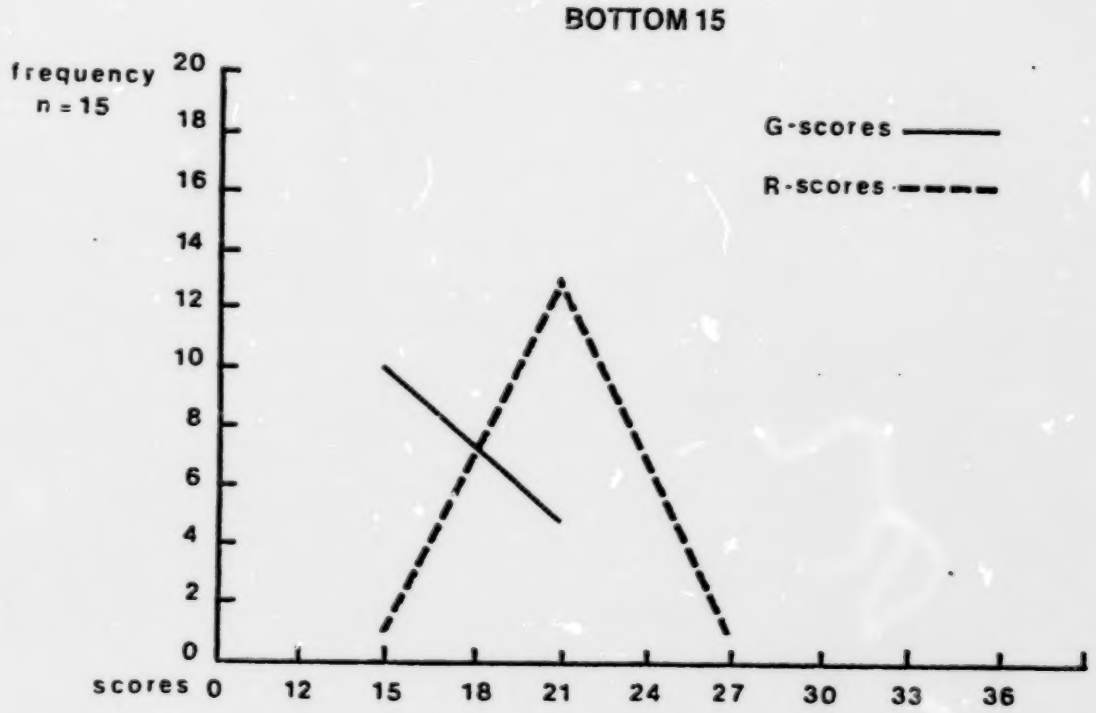


NOTE: Sums of scores: 2 topics, 3 raters, 6 point scale (possible range 6-36)

APPENDIX E

Figure 2: Frequency Distributions for Top and Bottom 15 Students' Scores





NOTE: Group means: 4 topics, 3 raters. 6 point scale (possible range 6-36)

APPENDIX F

Expressive Topics of Student 334

Topic E1

My mother thought piano lessons would be a good idea for me. She convinced me with ideas of fun, challanges, and culture. Being nine years old at the time, I believed her implicitly. What she didn't tell me, however, was that I'd have to take the grade one Royal Conservatory exam.

Piano lessons started off pleasantly enough. My piano teacher was fat,loud, and friendly and seemed to have been endowed with an unlimited amount of patience. Then one day he started assigning the same pieces over and over. When I questioned him about this, I was told that he was preparing me for my first exam. Suddenly he was no longer the kind, jolly man I once liked. He had turned into a tormenting tyrant who forced little girls to take frightening exams.

I practised and practised until I could play each song with my eyes closed. I learned about the octaves and the perfect fifths. I studied sight tests and ear tests until finally the dreaded day arrived.

My mother took me down on the bus to the building downtown. Inside, fifty other kids sat waiting for their turn with the examiner. I remember being amazed at the apparent calmness they exhibited as they chatted with their neighbors. As for me, my heart was racing fast and I barely noticed my surroundings as my turn came up.

I was led into a large completely white room. I felt like I was ruining its sterility. Suddenly a voice from across the room asked me to sit down at the piano. Unused to the piano, I was conscious of it's every squeak. What followed next, I have no clear memory of. All I know is that every piece I played, every instruction I obeyed was filled with flaws. When it was finally over, and I ran quickly to the comfort of my mother. As we left the relief and exhaustion that I was finished caused me to start crying.

I told everyone I had failed as I myself was convinced, so it was with much surprise and bewilderment that I read my exam results. I had received ninety-four percent.

Sum of Global Scores of three raters: 17

Sum of Rhetorical Effectiveness Scores of three raters: 18

Topic E2

I have been influenced by a great many people in my life and therefore it is impossible for me to pinpoint one as the most influential. In every school, job, vacation and in fact, every situation there have been people who, if only marginally, have contributed to the formation of my personality.

Parents have had much power in setting values for me. I have adopted many of their standards as well as their prejudices. Some of my strongest convictions and ideas were also those of my parents.

Teachers were certainly a strong influence over me, especially in my younger years. At that time I was not yet very opinionated and tended to believe everything I heard. Since then I've become considerably more discriminating but I'm quite impressed by much of what I'm told.

Like most teenagers I've been very influenced by my friends. When parents suddenly become the enemies we turn to our comrades for support and advice. We conform and dress and act alike until we realize that individualism is more important.

I expect that I shall continue to be influenced by the people and events I'm exposed to. It is impossible to change and grow in like without having first been influenced by someone.

Sum of Global scores by three raters: 9

Sum of Rhetorical Effectiveness scores of three raters: 11

Transactional Topics of Student 329

Topic T1

We are born equal. Traditional sex roles were a real fuss made up by the male chauvinists. I believe that women as as capable as men in doing their job or even better. Women are more responsible and understanding while men are aggressive and reckless in nature.

People used to say that a girl should stay at home and be pretty enough to marry a rich guy. These, in my own opinion, are all junks. We are born with the same mentality, quality and rank. For example: Queen Elizabeth I, she has all the quality of a ruler and at the sametime being a woman. The women came out to help during war world II and look at the business ladies, aren't they as capable as men? It's hard for some people to accept the fact that women are as capable as men because they are naturally brought up with the propagated idea that women are of lower rank and only best suited to be a housewife and have lot of babies.

I stated that women are more responsible and understanding because they really are. They can take care of children, work and at the same time be a housewife. As for men, they are really nothing but immature brats who only know how to fuss about when they get home and expect everything to be ready for them.

I believe that women are as good as men in that both of them have the potentiality to do what they are gifted in. Women in Edmonton or anywhere else do heavy, outdoor work for the department of public works are women who know what they are doing. As long as they are honest and sincere, and think that they have the capability, I don't see why people has to make such a fuss about it. There is rights and wrongs for everything but it is up to each and every individuals to make their own decision about it.

Sum of Global Scores of three raters: 6

Sum of Rhetorical Effectiveness Scores of three raters: 8

Topic T2

Many private clubs and organizations try to limit the membership of their organization on the basis of sex, ethnic group or religion. Exclusion of the opposite sex is common and based on clubs that provide services for only one sex. Ethnic groups may be excluded for reasons of ethnic differences between members of different ethnic groups. Religious groups may be excluded if a group preaches a religion that clashes with the religious beleifs of the club. Clubs should have the right to limit memberships on

these basis, as long as the club is of lawful intent.

Mens or womens clubs may exclude members of the opposite sex for several reasons. The club may provide activities for only one sex, and may only support facilities for only one sex. To admit members of the opposite sex would increase costs greatly. The club may also have been created as a meeting place where members of the same sex could meet and talk with others on a close social basis. If the meetings were mixed, it could cause tensions and destroy the atmosphere of the club. Membership may be limited on a sex basis due to the possible increased costs involved and also if unlimited membership would destroy the traditional atmosphere of the organization.

Ethnic groups may be excluded from clubs because their ethnic beliefs and traditions may clash with current club members. If a club is composed of a group with similar views on ethnic ideals, a club should not be forced to admit a member with ethnic ideals that clash with the rest of the club. Certain ethnic groups may also have long standing disagreements with other groups. These groups should not be allowed to mix if the organizations do not want to. Exclusion of ethnic groups should be allowed to keep the nature of the group homogenous.

Religious groups may be excluded on the basis that their religion disagrees with that of club members. A club with similar religious views should not be forced to admit members of other religions which do not agree with their religious beliefs. Many churches are exclusive for their religion and clubs should be permitted to the same.

Any club should have the right to limit their membership by sex, ethnic group, or religion as long as the group is lawful in intent. Admission of excluded members could cause increased costs, disagreements between members, and an overall loss of traditional atmosphere of the club. Private clubs should be allowed to limit membership, otherwise they are not really private clubs.

Sum of Global Scores of three raters: 16

Sum of Rhetorical Effectiveness Scores of three raters: 15

References

Anderson, J. The English placement test project. Vancouver: Educational Research Institute of British Columbia, 1979. (a)

Anderson, J. Personal communication, August, 1979. (b)

Appleby, B. N. Topic construction. Paper presented at the meeting of the National Council of Teachers of English, Seattle, April, 1983.

An assessment of the writing performance of high school seniors. Sacramento, CA: California State Department of Education, 1977.

Bailey, K. D. Methods of social research. New York: The Free Press, 1978.

Bereiter, C. Discourse type, schema, and strategy--A view from the standpoint of instructional design. Paper presented at the American Educational Research Association/International Reading Association Symposium, Toronto, April, 1978.

Bossone, R. & Larson, R. Needed research in the teaching of writing. New York: City University of New York, Center for Advanced Study in Education, 1980. (ERIC Document Reproduction Service No. ED 184 136)

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in written composition. Champaign, IL: National Council of Teachers of English, 1963.

Bradley, R. Statistical report. In E. M. White (Ed.). Comparison and contrast: The 1979 California State University and Colleges Freshman English Equivalency Examination. Los Angeles: Trustees of the California State University and Colleges, 1980.

Breland, H. M., & Gaynor, J. L. A comparison of direct and indirect assessment of writing skill. Journal of Educational Measurement, 1979, 16, 119-128.

British Columbia Post-Secondary English Placement Test.

Vancouver: Educational Research Institute of British Columbia, 1974-1983.

Britton, J. The composing processes and the functions of writing. In C. R. Cooper & L. Odell (Eds.). Research in Composing. Urbana, IL: National Council of Teachers of English, 1978.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. The development of writing abilities (11-18). London: Macmillan Education, 1975.

Britton, J., Martin, N., & Rosen, H. Multiple marking of English compositions. (Schools Council Examinations Bulletin No. 12). London: Her Majesty's Stationery Office, 1966.

Chase, C. J. The impact of some obvious variables on essay test scores. Journal of Educational Measurement, 1968, 5, 315-318.

Checketts, K. T. & Christensen, M. G. The validity of awarding credit by examination in English composition. Educational and Psychological Measurement, 1974, 34 (2), 357-361.

Coffman, W. E. On the validity of essay tests of achievement. Journal of Educational Measurement, 1966, 3 (2), 151-156.

Conry, R. F. & Jeroski, S. F. General report: British Columbia assessment of written expression. Victoria, BC: Ministry of Education, Learning Assessment Branch, 1980.

Conry, R. F. & Rodgers, D. R. Summary report: British Columbia assessment of written expression. Victoria, BC: Ministry of Education, Science, and Technology, Learning Assessment Branch, 1978.

Cook, T. D. & Campbell, D. T. Quasi-experimentation: Design and analysis for field settings. Boston: Houghton-Mifflin, 1979.

Cooper, C. R. Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.). Evaluating writing. Urbana, IL: National Council of Teachers of English, 1977.

Cooper, C. R., & Odell, L. Evaluating writing. Urbana, IL: National Council of Teachers of English, 1977.

Crowhurst, M. The effect of audience and mode of discourse on the syntactic complexity of the writing of sixth and tenth graders. Unpublished doctoral dissertation, University of Minnesota, 1977.

Crowhurst, M. Syntactic complexity and teachers' quality ratings of narration and arguments. Research in the Teaching of English, 1980, 14 (3), 223-231. (a)

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980, 5 (1), 6-13. (b)

Crowhurst, M. Sentence combining: Maintaining realistic expectations. College Communication and Composition, 1983, 34 (1), 62-72.

Crowhurst, M., & Piche, G. L. Audience and mode of discourse effects on the syntactic complexity of writing at two grade levels. Research in the Teaching of English, 1979, 13 (2), 101-110.

Deale, R. N. Assessment and testing in the secondary school. London: Evans/Methuen Educational, 1975.

Diederich, P. Measuring growth in English. Urbana, IL: National Council of Teachers of English, 1974.

Dixon, J. & Stratta, L. Achievements in writing at 16+. Birmingham, England: University of Birmingham, 1982. (a)

Dixon, J. & Stratta, L. Criteria for Writing in English. (Discussion booklet I: a report on the 1980 CEE Cross-Moderation Exercise). Southampton: Southern Regional Examinations Board, 1981.

Dixon, J. & Stratta, L. Teaching and assessing argument. (Discussion booklet 2: a report on the 1981 CEE English Cross-Moderation Exercise). Southampton: Southern Regional Examinations Board, 1982. (b)

Ebel, R. L. Essentials of educational measurement. Englewood

Cliffs, NJ: Prentice-Hall, 1979.

Egan, K. Educational development. Toronto: Oxford University Press, 1979.

Emig, J. The composing processes of twelfth graders. Urbana, IL: National Council of Teachers of English, 1971.

The English placement test. Vancouver: Educational Research Institute of British Columbia, 1982.

Evanechko, P. Language B. C.: An assessment in the English language arts (a pilot study). Victoria, BC: The Department of Education, 1976.

Evans, P. J. A. Evaluation of writing in Ontario: Grades 8, 12, and 13. Toronto: Ontario Ministry of Education, 1979.

Finn, P. J. Computer-aided description of mature word choices in writing. In C. R. Cooper & L. Odell (Eds.). Evaluating writing. Urbana, IL: National Council of Teachers of English, 1977.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures of grading English themes. Research in the Teaching of English, 1967, 1, 190-200.

Freedman, S. W. Influences on evaluators of expository essays: Beyond the text. Research in the Teaching of English, 1981, 15 (3), 245-255.

Freedman, S. W., & Robinson, W. S. Testing proficiency in writing at San Francisco State University. College Communication and Composition, 1982, 33 (4), 393-398.

Gilbert, P. I. F. The effects of a primary trait scoring guide on the reliability, validity, and time used in teacher evaluation of student writing. Unpublished doctoral dissertation, Ohio State University, 1980.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. Princeton, NJ: College Entrance Examination Board, 1966.

Greenberg, K. The effects of variations in essay questions on the writing performance of City University of New York freshmen. (Research Monograph Series No. 1) New York: City University of New York, 1981.

Guilford, J. P., & Fruchter, B. Fundamental statistics in psychology and education. (6th ed.) New York: McGraw-Hill, 1978.

Guth, H. P. The uses of language. Toronto: McGraw-Hill Ryerson, 1979.

Hansen, I. V. Evaluation in the secondary school. Paper presented at the meeting of the Canadian Council of Teachers of English, Montreal, May 12, 1983.

Hart, K. A., & Heim, A. C. Sentences, paragraphs, & essays. Toronto: Little Brown, 1979.

Hilgers, T. L. Experimental control and the writing stimulus: The problem of unequal familiarity with content. Research in the Teaching of English, 1982, 16 (3), 79-88.

Hiller, J. H., Marcotte, D. R., & Martin, T. Opinionation, vagueness, and specificity-distinctions: Essay traits measured by computer. American Educational Research Journal, 1969, 6, 271-286.

Hoetker, J. Essay examination topics. College Communication and Composition, 1982, 33 (4), 377-392.

Hopkins, K. D., & Stanley, J. C. Educational and psychological measurement and evaluation (6th ed.) Englewood Cliffs, NJ: Prentice-Hall, 1982.

Hunt, K. W. Grammatical structures written at three grade levels. Urbana, IL: National Council of Teachers of English, 1965.

Jeroski, S. F. Competence in written expression: Interactions between instruction and individual differences among junior high school students. Unpublished doctoral dissertation, University of British Columbia, 1982.

Kinneavy, J. L. A theory of discourse. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Knapp, T. R. Essay topics and modes and their effects on student prose. Paper presented at the meeting of the American Educational Research Association, Chicago, April 6, 1972.

Lloyd-Jones, R. Primary trait scoring. In C. R. Cooper & L. Odell (Eds.). Evaluating writing. Urbana, IL: National Council of Teachers of English, 1977.

McColly, W. What does educational research say about the judging of writing ability? Journal of Educational Research, 1970, 64 (4), 148-160.

McDaniel, B. A. The Chinese Canadian and the reliability of placement tests. Paper presented at the meeting of the Canadian Council of Teachers of English, Ottawa, May, 1979.

McMaster, R. J. Expression and communication. Don Mills, ON: Academic Press Canada, 1980.

Mellon, J. C. Transformational sentence combining. Urbana, IL: National Council of Teachers of English, 1969.

Mellon, J. C. National assessment and the teaching of English. Urbana, IL: National Council of Teachers of English, 1975.

Melton, V. S. & McCready, M. A. Survey of national trends in writing assessment. Ruston, LA: Louisiana Tech University, 1982.

Mullis, I. V. S. Highlights and trends in writing achievement, 1969-1979: Results from the National Assessment of Educational Progress. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April, 1981.

National Assessment of Educational Progress. Writing achievement, 1969-1979: Results for the third national writing assessment. (Vol. 1: 17 year olds) (Report No. 10-W-01). Denver: Author, 1980.

Newkirk, T. How competent are the writing competency tests? Paper presented at the Alabama Symposium on English and American Literature, Mobile, October, 1979. (ERIC Document Reproduction Service No. ED 179 958)

Page, E. B. The imminence of grading essays by computer. Phi Delta Kappan, 1966, 47, 238-243.

Page, E. B. Grading essays by computer: Progress report. In Proceedings of the 1966 invitational conference on testing problems. Princeton, NJ: Educational Testing Service, 1966.

Quellmalz, E. S., Capell, F. J., & Chou, C. Effects of discourse response mode on the measurement of writing competence. Journal of Educational Measurement, 1982, 19, 241-258.

Raymond, J. C. What we don't know about the evaluation of writing. College Communication and Composition, 1982, 33 (4), 399-403.

Rosen, H. An investigation of the effects of differentiated writing assignments on the performance in English composition of a selected group of 15/16 year olds. Unpublished Ph. D. thesis, London University Institute of Education, 1969.

Sanders, S. E., & Littlefield, J. H. Perhaps test essays can reflect significant improvement in freshman composition: Report on a successful attempt. Research in the Teaching of English, 1975, 9, 145-153.

Smith, M. J. Aspects of verbal behavior in written composition: Contrasts between formal and nonformal operational thinkers of college age. Unpublished doctoral dissertation, Rutgers, 1980. (Abstract)

Smith, V. H. Measuring teacher judgment in the evaluation of written composition. Research in the Teaching of English, 1969, 3, 181-195.

Spandel, V., & Stiggins, R. J. Direct measures of writing skill. Portland, OR: Northwest Regional Education Laboratory, 1980.

Steele, J. The assessment of writing proficiency via qualitative ratings of writing samples. Paper presented at the meeting of the National Council on Measurement in Education, 1979.

Thompson, R. F. Is test/retest a suitable reliability in most composition studies? Research in the Teaching of English, 1980, 14 (2), 154-156.

Ulin, R. O. Equivalent to what? GED high school equivalency tests. English Journal, 1982, 71 (1), 21-25.

University of British Columbia English composition test. Vancouver: University of British Columbia, Department of English, 1983.

Watson, C. The effects of maturity and discourse type on the written syntax of superior high school seniors and upper level college English majors. Unpublished doctoral dissertation, State University of New York at Buffalo, 1980.

White, E. M. Comparison and contrast: The 1973 California State University and Colleges freshman English equivalency examination. Los Angeles: Trustees of the California State University and Colleges, 1973.

White, E. M. Comparison and contrast: The 1979 California State University and Colleges freshman English equivalency examination. Los Angeles: Trustees of the California State University and Colleges, 1980.

Wilkinson, A., Barnsley, G., Hanna, P., & Swan, M.

Wilkinson, A., Barnsley, G., Hanna, P., & Swan, M. Assessing language development: The Crediton Project. Paper presented at the meeting of the Canadian Council of Teachers of English, Ottawa, May, 1979.

Wiseman, S., & Wrigley, J. Essay-reliability: the effect of choice of essay title. Educational and Psychological Measurement, 1958, 18, 129-138.

research in written composition. Research in the Teaching research in written composition. Research in the Teaching of English, 1980, 14 (1), 73-81.

Wrigley, J., Sparrow, F. H., & Inglis, F. C. Standards in C. S. E. and G. C. E.: English and mathematics. (Schools Council Working Paper No. 9) London: Her Majesty's Stationery Office, 1967.